Identification of Activation of Transcription Factors from Microarray Data

A Thesis

Submitted to the Faculty

of

Drexel University

by

Andrei Kossenkov

in partial fulfillment of the

requirements for the degree

of

Doctor of Philosophy

March 2007

DEDICATIONS

This work is dedicated my wife Olga who thinks that dedication in PhD thesis is a joke.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Identification of Activation of Transcription Factors from Microarray Data

Andrei Kossenkov

Aydin Tozeren, Ph.D.; Michael F. Ochs, Ph.D.

Signaling pathways play a critical role in cell survival and development by regulation of transcription factor activity causing necessary gene products to be produced in response to different stimuli. Although the task of detecting activities of signaling pathways is extremely difficult, recent advances in microarray technology promise progress in the field. There are many clustering and pattern recognition algorithms that have been applied to analysis of microarray data. However, these methods lack an ability to address the biological nature of the data and force assignment of one gene to a single co-expression group, while ignoring the fact that many individual genes are regulated by different signaling pathways in response to different stimuli, and therefore the genes should be assigned to multiple groups of co-expression. Another issue in microarray analysis is a low signal-to-noise ratio provided by the technology, yet most of the clustering methods do not even take errors of the measurements into consideration.

Bayesian Decomposition is an algorithm that decomposes microarray data into a set of biologically meaningful expression patterns that could be linked to certain signaling pathways and groups of genes that contain these patterns, allowing assignment of one gene to multiple patterns of expression. To address the problem of low signal-to-noise we modified the Bayesian Decomposition algorithm to allow inclusion of prior gene coregulation information to improve statistical power. We also created the Automated Sequence Annotation Pipeline to provide microarray data mining processes with annotation information at all steps and particularly to deduce the coregulation information for a given set of genes from transcription factor database TRANSFAC.

We validated enhancements done to Bayesian Decomposition on simulated and real biological data and showed that using coregulation information can improve ability of the method to recover correct results. The designed data mining process that uses the Automated Sequence Annotation Pipeline and the modified Bayesian Decomposition was applied to determine transcription factor activities linked to patient outcome in gastrointestinal stromal tumor (GIST) patients undergoing treatment with imatinib mesylate (IM, Gleevec). The study demonstrates genes that can be potentially used as biomarkers to predict GIST patient response to Gleevec treatment and activity of transcription factors that can contribute to difference in the response.

CHAPTER 1: INTRODUCTION

1.1 Cancer biology

Cancer is a leading cause of death around the world, resulting in over 6 million deaths per year. Cancer is caused by a series of events that transform a normal cell into abnormal, so that it grows and divides uncontrollably. There are three defined steps of the transformation process: initiation, promotion and progression [1] (Figure 1). The initiation step is characterized by DNA mutation due to various factors such as viruses [2] or environmental damage (air pollution [3], tobacco consumption [4], radiation [5], chemical carcinogens [6], etc.). Promotion refers to process of growth and proliferation of the initial mutated cell that results in a small tumor, which in the progression stage undergoes further growth and invades the tissue of origin, accompanied by morphological changes. Sometimes progression develops into metastasis, when cancer cells acquire an ability to travel through the blood stream to other tissues of organism.

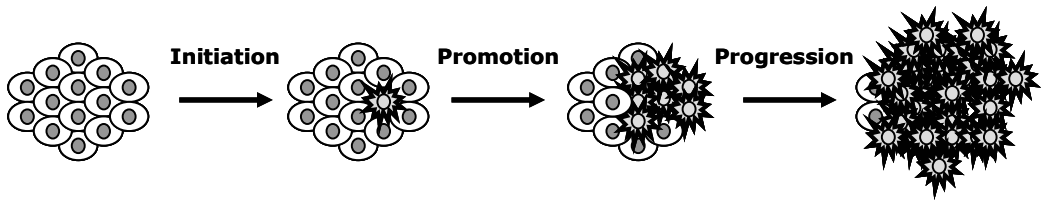

Figure 1. Initiation, Promotion and Progression stages of cancer development. DNA damage of one of normal cells (Initiation) result in uncontrolled growth and proliferation of the initial cell (Promotion) with following further growth and invasion of tissue of origin.

Upon detection, a cancer phenotype reflects mutations that occurred at the initiation stage and mutations that accumulated during abnormal cell development.

When coupled with genetic heterogeneity between patients, the result is a disease of great complexity and heterogeneity that produces different tumor progression rates, drug responses, and outcomes across patients. Therefore, one of the most important tasks in cancer research is to identify the specific genetic variations for an individual cancer that lead to the specific cancer phenotype. Although very heterogeneous at later stages, resulting in expression level changes of thousands of genes as documented by multiple studies of transcriptional profiling [7-10], cancer may be the result of very few genetic changes [11, 12], giving opportunity to understand precise mechanisms of the disease.

1.2 Signalling pathways and Cancer

In order to survive and develop normally, cells must react properly to various changes in their environment and their internal state. Varying environmental conditions will result in different cell reactions that require activity of specific proteins to perform necessary functions to adapt the cell to these conditions. One of the ways to get required proteins is to transduce a signal received from the environment through a pathway to activate transcription factors that initiate transcription of target genes. Such signaling pathways are highly controlled, so that a cell can provide a precise response to a certain stimulus. At the same time there are multiple pathways activated at the same time and connected with each other forming signaling networks. This makes a cell a very complicated system with elaborate control and repair mechanisms evolved to be stable, although sometimes because of accumulated mutations or under extreme conditions, signaling pathways may loose

appropriate regulation, which can lead to cancer, diabetes and many other disease states.

One of the most studied intracellular signaling pathways are the phosphorylation cascades that transduce signal through activation of mitogen-activated protein kinases (MAPKs). MAPK cascades play a key role in various important cellular processes, including response to external stimuli, growth, proliferation, differentiation and apoptosis. There are three major MAPK families in eukaryotic cells, including extracellular-signal regulated kinase (ERK, also known as p42/44 MAP kinase), JUN N-terminal kinase (JNK, also known as SAPK1) and p38 (also known as SAPK2) [13-16]. More specifically, ERK cascade is known to regulate cellular proliferation, differentiation, and survival, JNK pathway is related to stress response and apoptosis, and p38 also mediates response to environmental stress and is involved in other fundamental biological processes.

MAPKs cascades are activated by a variety of receptors leading to signal transduction through intermediate proteins and represent a complex network of consequent protein activations where different parts of the network are connected between each other to keep the system in balance. For example, crosstalk between RAS/RAF/MEK/ERK proliferation cascade and PI3K/AKT apoptosis related pathway represent fine-tuned balance to keep a cell under control [17].

Over the last few decades numerous signal transduction pathways have been reported whose dysregulation may play an important role in the growth and survival of cancer cells [18-22]. Signaling cascades can be distorted at different levels, starting from constitutive activation of tyrosine kinase receptors, further downstream

mutations of signaling genes or mutations of transcription factors, resulting in uncontrolled changes of transcription of genes that involve cell division, cell growth and survival that can lead to tumor development. For example, epidermal growth factor receptors (EGFR) activate survival signaling pathways including PI3K/AKT, RAS/RAF/MEK/ERK, and JAK/STAT signaling pathways (Figure 2). A number of cancers including breast and brain tumors [20, 23] show overexpression of epidermal growth factor receptors, causing undesirable cell proliferation as a response to smaller amounts of epidermal growth factor. Mutations in Ras proteins can imitate growth-promoting guanidine triphosphate (GTP)-bound Ras leading to downstream activation of the MAPK cascade leading to cell proliferation [24].

Pathway dysregulation can be determined by measuring gene expressions changes and recent advances in microarray technology allow measurement of mRNA expression levels for thousands genes simultaneously [25, 26]. Due to complicated processes of transcription, translation and protein activation, it is impossible to use these measurements of mRNA levels as upstream indicators of protein activity [27, 28]. Specifically, study of a subset of genes expressed in the *Saccharomyces cerevisiae* showed a low correlation of 0.356 for 73 selected genes between mRNA and protein expression levels [28]. While true for the yeast species, the complexity of mammalian systems contribute even more to the difference. Protein activation is yet another stage of regulation process that determine amount of functioning proteins in a cell and that is especially true for signaling proteins, which require post-translational modification to become active. As such, gene expression can be treated only as a downstream indicator of pathway activity.

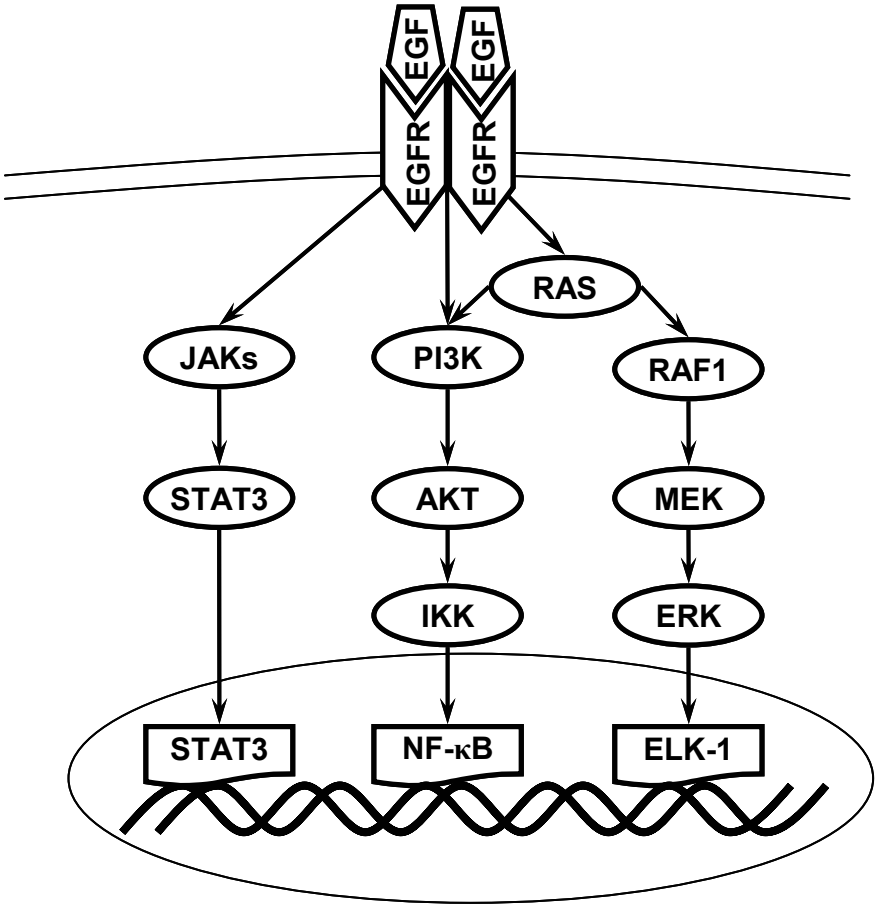

Figure 2. Part of epidermal growth factor (EGF) pathway.
EGF receptor (EGFR) transduces signal from EGF to RAS/RAF/MEK/ERK, PI3K/AKT and JAK/STAT pathways, that activate transcription factors related to cell survival and proliferation. Transcription factor STAT3 induces progression through the cell cycle and prevents apoptosis, NF-κB – regulates expression of anti-apoptotic genes, and ELK-1 – mediates growth factor stimulation of proliferative response.

1.3 Gastrointestinal Stromal Tumors

Gastrointestinal stromal tumors (GISTs) are rare but deadly mesenchymal tumors affecting approximately 2,000 people in the U.S. per year. GISTs are the most common gastrointestinal mesenchymal malignancies and may occur anywhere along the gastrointestinal tract, but most often occur in the stomach and small

intestine, representing 60%- 70% and 20%-30% of tumors, respectively [29]. GISTs are discovered either incidentally during endoscopic, radiologic, or surgical procedures; or are diagnosed in the evaluation of patients with an abdominal mass, abdominal pain, or upper gastrointestinal bleeding. Complete surgical resection is still the only treatment that can completely cure the disease. However, even for patients whose tumors are fully removed 5-year overall survival is only 50%, use of chemotherapy and local radiotherapy appeared ineffective in patients with unresectable or metastatic GISTs, which represent 30% of tumors, giving a median survival ranging from 9 to 20 months [30, 31].

GISTs are considered to originate from the interstitial cells of Cajal (ICC) [31, 32], since they share many of the phenotypic features. Experiments show that 90%-95% of GISTs have c-KIT gain-of-function mutations, mostly in exon 11 but also in exon 9,13, or 17 [33, 34], while the other cases have gain-of-function mutations in the platelet-derived growth factor receptor alpha [35]. The mutations of KIT occur somatically and lead to constitutive, ligand independent activation of KIT and its signal transduction pathways, such as the PI3K/AKT and RAS/RAF/MEK/ERK pathways, and probably the STAT3 pathway [36].

Gleevec (imatinib mesylate, STI-571; Novartis, Basel, Switzerland), a derivative of 2-phenylaminopyrimidine, is a small molecule with activity against a number of related protein tyrosine kinases, including KIT, PDGFR, ABL and BCR-ABL [37-40]. In a nation-wide phase II clinical trial more than 81% of patients with unresectable or metastatic GIST benefited from Gleevec therapy; 53.7% had a partial response and 27.9% achieved stable disease [41].

While a significant success in treatments of GISTs with Gleevec was shown lately, there is still a little understanding of cases with no or poor response to the treatment. Some mechanisms suggested for such resistance included secondary point mutations of KIT or PDGFR [42, 43] or activation of alternate receptor tyrosine kinase protein [44], but overall picture is still unclear and requires additional research efforts. An ongoing multi-institutional clinical trial performed through the cooperative group of Dana-Farber Cancer Institute, Oregon Health Sciences University, Fox Chase Cancer Center and University Hospital of Helsinki focuses on the problem and is generating microarray measurements on biopsies from GIST patients before and surgical samples after Gleevec treatment. We have applied a modified version of the Bayesian Decomposition algorithm [45] to analyze the data in order to link patients' response to treatment to changes in signaling pathways and understand conditions leading to non-response to Gleevec.

1.4 Microarrays

First high-throughput gene expression microarrays used radioactively labelled targets hybridized onto cDNA probes grown on membrane-based arrays [46]. This then developed into microarrays that use fluorescent labeling to avoid problems with stability, handling, disposal and safety risk of radioactive compounds. First introduced in 1995 [26], hybridization of fluorescently labeled targets to cDNA microarrays printed on glass targeted 46 cDNA probes simultaneously and already by the end of 1996 cDNA chips with 1,000 probes were reported to be used in experiments [47, 48]. Since then rapid growth in the technology allowed millions of

probes to fit on a 1.28 cm^2 chip (CeneChip expression arrays from Affymetrix), enough to cover whole genome of an organism.

While various microarray systems use different chip printing processes, the two most important chip types are one-channel arrays (Affymetrix) – *in situ* synthesized oligonucleotide arrays that use single fluorescence channel to measure expression level of genes of a sample that build up oligos directly on a slide, and two-channel arrays (spotted arrays) made by depositing pre-made oligos or cDNAs onto slides and two fluorescence are used to label experiment and control samples before hybridization.

In order to measure expression levels of genes, messenger RNA (mRNA) from sample of interest is extracted, converted to complimentary DNA (cDNA) or complimentary RNA (cRNA) and tagged with a fluorescence label that can be detected with a scanning device at a certain wavelength. For two-channel arrays reference mRNA is also required to be prepared and tagged with a different fluorescence label. Then dyed products are hybridized on a microarray slide, excess labelled oligonucleotides are washed off and the microarray is scanned. The whole experimental process is schematically shown in

Figure 3. Although these two approaches are very different in design, recent studies suggest that both methods are essentially equivalent when compared on controlled data [49].

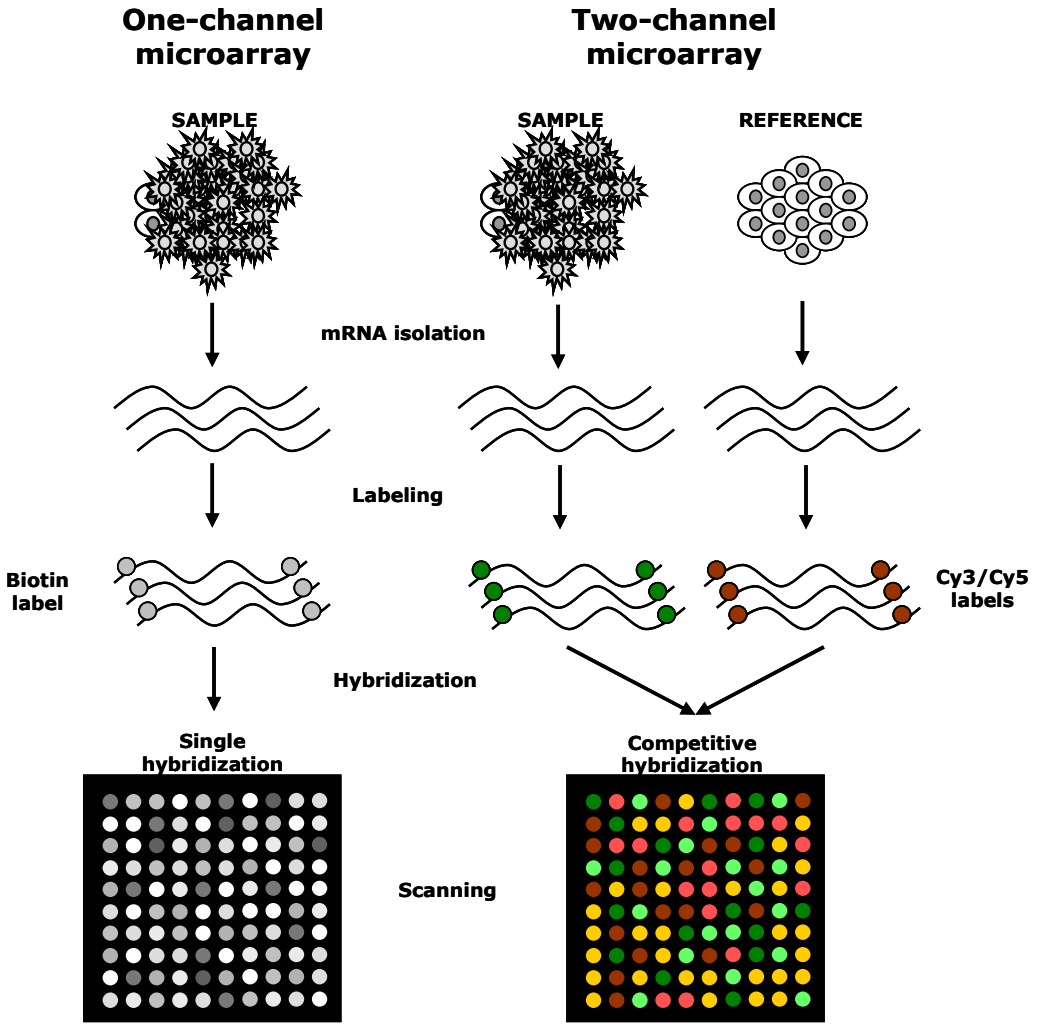

Figure 3. Schematics of experimental process using one-channel and two-channel microarrays.

Microarray technology is a very powerful tool for molecular biology, but it has certain drawbacks that limit the effectiveness of microarray experiments. First, biological variations, artifacts of preparing samples for analysis, inconsistencies of array chip printings, all contribute to low signal-to-noise in microarray technology, which frequently results in inconsistent results. The impact from the problem can be

lowered by performing replicate experiments and using analysis methods that can take into account the estimated noise of the system. Another issue, as mentioned earlier, is that measurements of mRNA levels can not be used as a direct indicator of corresponding protein activity due to complicated multi-step post-translational modifications. Although it limits the interpretation of microarray results, it is still possible to use changes in expression levels as downstream indicators of biological processes of the cell.

Correct gene assignments and other related annotations are also of great importance for analysis and result interpretation of microarray data. Improper probe annotations can be a result of various uncertainties. Mismatches of clones from cDNA libraries used for spotted microarray construction can produce a significant amount of errors in identifications for these probes. In fact, a study of commercial subset of the IMAGE Consortium mouse cDNA clone collection showed that only 62.2% of 1189 analyzed stocks were correctly identified [50]. Another example is a work of Harbig *et al.* who re-identified the probesets on the Affymetrix U133 plus 2.0 GeneChip array that resulted in redefinition of approximately 37% of the probes [51]. Non-specificity of probes to splice variants and overlapping transcripts that is sometimes ignored due to limited knowledge can be another reason for annotation error. Also, there are many mistakes in clustering of Expressed Sequence Tags (EST), which frequently were used in older platforms to design microarray probes to target specific genes, resulting in false assignment of EST to certain gene. More recent arrays are designed against the genome, but annotations of the GenBank genomic accession numbers are imperfect, and although many errors are corrected constantly,

it is clear that there are still inconsistencies in current assignments of microarray probes that can significantly affect results of experiment. Therefore, an ability to acquire the most recent annotation information is crucial for receiving correct results.

1.5 Microarray data analysis

Prior to microarray data analysis, several pre-processing steps are required. First, scanned images are quantified to determine the signal intensity of each spot. This is usually done by software specifically designed for a specific microarray platform that the experiment was performed on, although there are some other alternative quantification methods and their modifications available, especially for spotted arrays, e.g. GenePix [52], ImaGene [53], TIGR Spotfinder [54], WaveRead [55], and many more [56]. Second, normalization of microarray data is performed to remove dye-related differences between two channels and various slide-specific artifacts that can exist between different microarray. After pre-processing, the normalized microarray data can be analyzed using statistical methods, clustering techniques or more advanced pattern recognition approaches.

1.5.1. Normalization

When performing experiments with multiple microarray slides, there are always sources of non-biological variation between arrays such as dye biases, sample preparation or hybridization differences, scanner calibrations, slide printing variations, volume of initial RNA, etc. To correct some of this variability, a series of steps referred to as “data normalization” are performed on the data before analysis. While the main assumption behind normalization of microarray data is that most of

the genes on the slide do not change their expression levels and numbers of up- and down-regulated genes on the array are roughly equal, most methods try to adjust expression levels of the genes, so overall average expression remains the same across different arrays. Additional steps can include removing saturated signals from microarray, background correction, low expression genes correction, etc.

For two-channel microarrays, many different methods have been developed in order to compensate for dye-effects and other systematic errors between arrays. Total intensity normalization [57], for example, transforms expression ratios between channels in such a way that mean log2(ratio) across all measurements of the array is equal to zero. There are other methods that use similar global normalization approach, including log centering, rank invariant methods [58], and many other variations. While such methods do not take into account situations when dye-effects depend on signal intensity and/or spatial location within the array [59], locally weighted linear regression (LOWESS) method [60] accounts for such effects and has been proven to be a robust, powerful normalization method for different types of two color microarray experiments [59]. Although many new methods [61-64] and modifications [65, 66] have been proposed and compared to LOWESS, the comparison results are inconsistent and new methods outperform LOWESS only in special cases [67].

Affymetrix arrays, designed to measure abundance of mRNA levels using only one channel, require different approaches to normalize the data. There are several widely used methods developed for expression data. Microarray Suite (MAS) from Affymetrix uses a linear regression method for perfect match (PM) values [68].

Introduced later, dChip software uses a model-based expression index (MBEI), with a chip showing median intensity selected as a baseline array, an invariant set of probes used for comparison between two samples, and a non-parametric curve (running median) is fitted through the data points [69]. The most recently developed method, robust multi chip average (RMA) use quantile normalization, where highest PM intensities (background corrected and log transformed) are replaced by their average and the process is repeated for all intensities in descending order [70]. A modified version of RMA, GCRMA has been developed to account for GC rich probes having higher intensities due to increased binding, and models probe intensity as a function of GC content of the probe [71].

1.5.2. Statistical methods

Fold change (FC) was one of first methods used to identify genes that are differentially regulated between two conditions (samples, time points, etc.) by selecting genes with fold change (ratio of the measured response in one condition to that in another) that is outside of a given cutoff. Although very popular from when microarrays were first introduced due to its simplicity, the method has major drawbacks and has been called into question [72-75]. The inability to incorporate variance, provide confidence intervals for results, and potential high false discovery rate (FDR) of the method resulted in attempts to develop and apply more stringent statistical approaches.

In order to assign some level of confidence to inference that a gene is differentially expressed between conditions, various statistical methods were

suggested for microarray analysis, among which Student's t-test, Mann-Whitney-Wilcoxon rank test, analysis of variance (ANOVA) and significance analysis of microarray (SAM) are among most popular methods. ANOVA uses Fisher's F-distribution as part of the test of statistical significance and compares group variations to the overall variation observed [76]. There are variations of ANOVA analysis, including one-way, factorial or non-parametric ANOVA, that are used depending on experimental design or hypothesis an investigator wants to test [77, 78]. Student's t-test is used to test the hypothesis that a gene's expression levels differ between two sets of samples by using the T statistic and determining the significance level of the difference from *t* distribution [79]. SAM test uses slightly different statistic that is based on t-statistic, but also uses a correction [80] that reduces the relative differences for low expressed genes and genes with similar expression levels. A permutation procedure is also used to estimate the false discovery rate for the final results [80]. Mann-Whitney-Wilcoxon rank test is a non-parametric statistical test, that like a t-test and SAM, compares for each gene the difference between measurements in two groups. However, it does not require assumptions about the form of the distributions of the measurements, so it is more reliable when used on microarray data with large number of outliers or high noise. The method's statistical power strictly depends on sample sizes, and provides poor significance levels for groups with fewer than 6 samples.

Using simple statistical approaches to test thousands transcripts in one experiments is likely to result in many false positive results with significant confidence level. While the significance of confidence levels are usually depend of

the statistical test performed on the data, a proper adjustments are required to compensate for large number of tested genes or for small number of samples [81, 82]. On the other hand, experiments with a low number of samples usually do not have enough statistical power to produce significant results. While performing tests based on gene-by-gene calculations is inefficient, it is possible to 'borrow' information across genes from microarrays to improve statistical power of the results, as clustering and pattern recognition methods do when combine genes together based on their expression profiles.

1.5.3. Cluster analysis

Cells have evolved to survive by reacting to different internal and external environmental conditions with a response that results in activating a set of proteins required to use or oppose these conditions. To optimize the process, genes whose products function together are usually undergoing same regulatory mechanisms so they are coordinately expressed in response to stimuli. This property is used by many clustering methods that group genes together based on their expression profiles and associate such groups of transcripts with a biological function or biological process that they are involved in. In this case microarray data for analysis with clustering methods can be represented by a matrix with measurements of genes (rows) for multiple conditions (columns), where conditions can be of various kinds of samples, e.g. different treatments, time points, patients, etc.

There are many widely used clustering algorithms for analysis of microarray data, including hierarchical clustering [83], quality threshold clustering [84], *k*-means

clustering [85], and self-organizing maps [86], among many other methods reviewed elsewhere [87, 88]. These methods differ from each other considerably and lead often to different results, even within the same method when using different distance metric as a measure of similarity between genes.

Clustering is a common technique widely used in many fields and various methods were borrowed and adopted for microarray analysis. Hierarchical clustering is one of the first clustering algorithms applied to microarray data [83, 89-92]. Using a distance metric (Eisen's original metric [83]), the method builds a hierarchical binary tree (called a dendrogram), starting from the individual genes' expression profiles as leaves (also thought of as separate clusters) by progressively merging clusters, where each internal node represents the average of its two children. The constructed tree can be cut at some point according to a threshold value to receive clusters of required characteristics. Due to its simplicity and clear representation, hierarchical clustering has been used in many reported microarray experiments, but a number of drawbacks should be considered. First, hierarchical clustering is a greedy search algorithm, meaning that merging decisions on early steps are based only on the distance between nodes and cannot be undone, but not necessarily the best ones in global scale and can lead to mistakes in the overall clustering. Second, dendrograms and corresponding heatmaps, which used extensively in visualizations of the analysis results, suffer from inversion problems that complicate interpretation of the hierarchy [93]. In addition, complexity of dendrograms for larger data sets makes them difficult to understand, and the choice of location for tree cut to receive final clusters is unclear. And finally, analysis of yeast cell-cycle dataset with hierarchical clustering

performed by Cherepinsky *et al*. showed that the method has very low accuracy of gene assignments to clusters (less than 60%) [94].

The *k*-means clustering starts from randomly dividing genes into *k* groups and calculating cluster centers (or centroids) for each of these groups. New groups are formed by reassigning each gene to the closest centroid. Then the centroids are recalculated for the new clusters and the process repeats [95]. While simplicity and speed of the method are the main advantages of the method, the most important disadvantage is that results are not unique across different runs and depend on starting positions of centroids. Another obstacle for analysis of microarray data is unknown number of clusters prior to analysis that needs to be estimated somehow.

The quality threshold clustering algorithm (QT Clust) is more computationally intensive than hierarchical or *k*-means clustering, but does not require specifying the number of clusters prior to analysis and always returns the same result for each run [84]. The algorithm iterates as follows. A maximum diameter for clusters is chosen before the analysis, and a candidate cluster is formed by first gene data point. Other data points are iteratively added by including the closest, based on jackknife correlation [84], point until no data points can be added to the cluster without surpassing the diameter threshold. Next candidate cluster is formed by second gene data point and all other data points, including those from first cluster, are considered for the cluster. The process repeats for all genes of the analysis. At first step, the number of clusters equal to the number of genes, and the largest candidate cluster is set as a real cluster, whereas all the genes from this cluster are removed for further analysis. The method recurses by analyzing remaining set of genes. As

mentioned earlier, the method does not require a number of clusters prior to analysis. However, a threshold for cluster diameter is required and hard to estimate. Although there was an attempt to improve algorithm by estimating appropriate threshold from the data itself [96], the problem of setting different diameter threshold for different clusters is not resolved.

The method of self-organizing maps (SOM) performs the following procedure. After choosing an initial grid of nodes (usually one- or two-dimensional), the nodes are mapped randomly into k-dimensional space. At each step of the algorithm a random data point is chosen and nodes are moved in the direction of it. Nodes are moved depending on distance between a node and that data point, so the closest node is moved the most compared to more distant nodes. After a number of such iterations, nodes represent clusters, with neighbor nodes in initial grids defining related clusters. Although a number of successful application of SOMs have been reported [86, 97-100], several disadvantages exist for the method. Sensitivity of SOM to incomplete data is a problem that is very important in microarray data analysis, due to abundance of missing data points resulting from data flagging during preprocessing. Also, a SOM can yield different decompositions of the data depending on the choice of initial conditions. Another issue is that initial node grid is fixed and may not be changed during the analysis, and that can lead to inappropriate mapping of the data space. In addition, when two data points are mapped from high dimensional space to nearby locations on the two dimensional grid, it is possible that those points are actually far apart in the higher dimensional space.

These methods have been used for the majority of published microarray studies, but all of them have a disadvantage of forcing one gene into single co-expression cluster, when many individual genes are involved in more than one process of the cell and therefore co-express in multiple groups [101]. For example, a study of yeast cell cycle microarray data showed that only about 10% analyzed genes were mapped to a single cell cycle phase, while the rest were regulated in multiple phases [102]. Some attempts were done to alleviate the problem, for example in fuzzy *k*-mean clustering [103] that uses principal component analysis to identify overlapping groups of objects by allowing the objects to belong to more than one group.

1.5.4. Advanced methods

More advanced, pattern recognition algorithms in contrast to clustering methods try to recover a set of underlying patterns of expressions that combine differently for each gene to result in the observed gene expression profile. In this case, the expression profile of each gene can be interpreted as a mixture of different expression patterns that relate to different biological processes, therefore, allowing one gene to belong to multiple groups of co-expression. Mathematically, pattern recovery maps to a process of finding such matrices, the product of which equals the original microarray data matrix plus the noise. Examples of matrix factorization approaches are Singular Value Decomposition (SVD), Principal Component Analysis (PCA, special case of SVD), Non-negative Matrix Factorization (NMF), Independent

Component Analysis (ICA), and Bayesian Decomposition (BD) (these methods are reviewed in Chapter 2).

1.6 Gene Annotations

Microarray data analysis is a very demanding task involving multiple steps of data pre-processing, normalization, filtering, data mining, and interpretation of results. Many stages of the process can be enhanced by additional biological knowledge about the objects being analyzed, e.g. the gene linked to each probe, groups of co-expressed genes, or gene ontology (GO) information, which has recently become very popular [104-107] to aid validation of clusters or expression patterns recovered by data analysis. Therefore, comprehensive and structured annotations for all probes on a microarray slide are very essential for the analysis. Currently, most microarray platforms provide annotation files for each chip they produce, however, these annotations are usually partial, outdated and unsuitable for high-throughput experiments. While it is possible to perform manual annotations for a limited number of probes using external annotation databases, such a non-automated process is very time consuming and requires deep knowledge of information sources to be able to combine collected information.

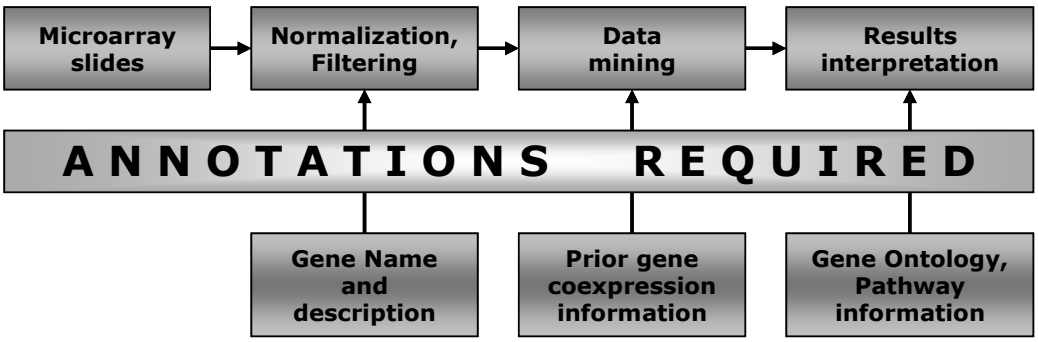

Figure 4. Example of annotations required at each step of microarray data analysis

Figure 4 describes basic microarray data processing steps and types of annotation data that can be utilized at each of these steps. Generally, normalization does not require any biological knowledge about data points being normalized, but there are methods that use housekeeping or other constitutively expressed genes for normalization of data. Combining replicates can be another step prior to data mining and depending on the level of replicates it requires mappings of a probe to sequence accession number or gene symbol. Filtering is yet another pre-processing step that uses gene annotations of probes. Apart from removing entries that do not vary over experimental conditions or do not have specific expression profile determined by the design of the experiment, the process can also be aimed at filtering out probes that do not have any gene associated with them in order to focus analysis on true signals that are actually belong to known genes.

Depending on data mining algorithm used for analysis of microarray data, gene annotations can be used to form training sets from genes of specific functional class for supervised algorithms, e.g. support vector machines (SVM) [108, 109] or neural

networks [110]. Another type of prior information for the analysis can be in a form of genes grouped by the biological process or by shared transcription factor. Such additional information added to the analysis can benefit the efficiency of an algorithm or improve statistical significance of final results. This prior knowledge can also be used for the algorithm validation when the method results are compared to already known information.

Results interpretation step can use a very broad variety of annotation information about genes under study, including gene ontology information, pathways, transcription factors, chromosomal location, etc. Analysis of common features between genes from the same groups determined by data mining can help to link these features to observed expression behavior. For example, abundance of genes related to apoptosis in a group can be linked to response of cancer patients to a treatment based on expression pattern common for these genes, whereas a common transcription factor for these genes can indicate activity of a certain upstream signalling pathway that lead to such response.

Several systems have been proposed and implemented to provide integrated access to a number of various available genomic databanks. There are different types of such systems that differ in data management: data warehousing, for example NCBI/Entrez [111], Swiss-Prot/TrEMBL [112], Ensembl [113], Kyoto Encyclopedia of Genes and Genomes (KEGG) [114], Gene Ontology Consortium [115], which host and manage unique information. Information linking systems, e.g. the most popular SOURCE [116], GeneLynx [117], GeneCard [118], mediate access to different data sources through a Web interface to receive integrated annotations that can be linked

to original sources. However, most of these systems provide only one gene at a time annotations and are designed for human access, but not for computer-based querying. Thus, these systems are limited for use when thousands of diverse annotations are required, especially when results from one search should be used as an input for different data source. Therefore, large datasets require additional automated tools for gathering and composing the information of interest from available databases.

CHAPTER 2: PATTERN RECOGNITION METHODS OF MICROARRAY ANALYSIS

2.1 Singular Value Decomposition and Principal Component Analysis

The Singular Value Decomposition (SVD) and closely related Principal Component Analysis (PCA) method in application to gene expression data were first introduced by Alter *et al.* [119], where they analyzed the yeast cell cycle data set generated by Spellman *et al.* [120]. Later, these methods became the most popular matrix factorization algorithms and have been reported to be successfully applied to many other datasets, for example genetic profiling in leprosy [121], analysis of Down syndrome [122], human fibroblast data [123], breast tumor classifications [124], tissue specific gene expression pattern search [125], to name a few.

The basic concept behind the SVD is the following. Let ***D*** *[n, m]* denote a data matrix of *n* genes over m samples (conditions) with rank *r* as shown in Figure 5. In this case d_{ij} is the expression level of the i^{th} gene in the j^{th} sample. The elements of the i^{th} row of ***D*** form the *m*-dimensional vector g_i, which is referred to as the transcriptional response or expression profile of the i^{th} gene. Alternatively, the elements of the j^{th} column of ***D*** form the *n*-dimensional vector a_j, which is referred to as the expression profile of the j^{th} sample. Singular value decomposition of the matrix ***D*** produces two orthonormal bases, one defined by right singular vectors and the other by left singular vectors, as described by the equation:

$$D = USV^T \tag{1}$$

where columns of ***U*** *[n, m]* are the left singular vectors (form an orthonormal basis for the sample expression profiles), ***S*** *[m, m]* is a diagonal matrix of ordered singular values, and the rows of $\boldsymbol{V^T}$ *[m, m]* are the right singular vectors corresponding to ordered singular values that form an orthonormal basis for the gene transcriptional responses (Figure 5). Therefore, gene transcriptional response g_i can be described as a linear combination of the right singular vectors also called eigengenes. Alternatively, sample expression profile a_j can be presented as linear combination of the left singular vectors called eigenassays. In other words, eigenenes represent expression patterns found in the data and eigenassays define what genes contain corresponding pattern.

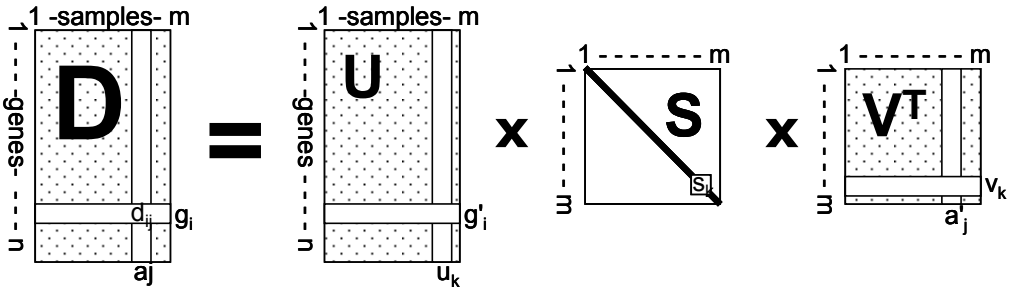

Figure 5. Singular Value Decomposition. Initial matrix ***D*** is decomposed into product of left singular matrix ***U***, diagonal matrix of ordered singular values ***S***, and right singular matrix $\boldsymbol{V^T}$

In some cases it may be practical to reduce matrices dimensionality to $p < m$, then only the p largest singular values are calculated, while the rest of the matrix is discarded (Figure 6). This way, only p expression patterns will be found in the process. This approach is much quicker and more economical than SVD for $m >> p$

columns. The truncated SVD is not an exact decomposition of the original data matrix, however, the approximation may be sufficient for practical applications, especially to remove signals that represent the noise.

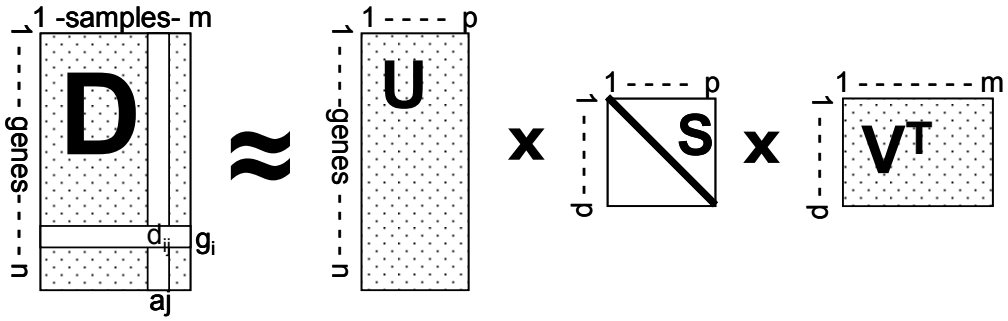

Figure 6. Truncated Singular Value Decomposition. It is possible to discard smaller singular values, keeping only first p singular values that keep most of the expression information.

Principal component analysis is sometimes used as a synonym to SVD and is actually a special case of singular value decomposition, i.e. PCA uses SVD to project initial matrix into reduced space. PCA projects data into direction with the most data variance via linear transformation. A new coordinate system is selected in such a way that the greatest variance of the data is located on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on, while principal components are orthogonal to each other. Retaining of only $p < m$ lower-order principal components as expression patterns allows matrix dimensionality reduction while keeping the strongest data variations.

One of the most valuable features of SVD and PCA is that it is possible to determine a number of expression patterns that explain the data by filtering out the

eigengenes (or high-rank principal components in case of PCA) that represent noise or experimental artifacts [119]. Using this property, PCA can be used to obtain true dimensionality of the data for methods that require to define number of clusters prior to analysis [126]. Also, SVD is capable to detect biologically meaningful patterns of expression, even when clustering methods fail due to weak signals in the data [127].

The major issue of PCA method is the restriction of orthogonality it imposes on underlying expression patterns. If the observed signals (PCs) are not orthogonal by nature, which is generally true for biological data, PCA does not produce biologically meaningful expression patterns. Another SVD/PCA shortcoming is that they do not take into account any error measures associated with the data points, although a recent modification made for PCA tries to solve the issue [128]. Absence of parameters to tweak and coefficients to adjust for the method can also be considered as an issue for the method, because some prior information about observed signals is available, an inability to incorporate this information into analysis is a significant drawback of the method.

2.2 Independent Component Analysis

Application of the Independent Component Analysis (ICA) approach to gene expression data was introduced by Liebermeister [129]. He compared ICA with PCA to show that introduction of non-orthogonal basis for dimensionality reduction is more biologically meaningful and takes into account high-order dependancies in the data. The original work [129] analyzed yeast cell-cycle data [120] and B-cell lymphoma data [130]. In more recent studies, ICA has also been applied to

classification of ovarian cancer [131], study of endometrial cancer [132] and diagnosis of human cancer types [133].

Like PCA, ICA performs data matrix decomposition by projecting initial data on a lower dimensionality space. However, by removing all linear correlations, ICA allows a non-orthogonal basis for such decomposition, but it still requires statistical independence of components between each other. By stating that observed microarray signals are a result of a mixture of underlying biological processes in the cell, decomposition of matrix $\boldsymbol{D}$ can be expressed by the following equation:

$$D = f(AP) \tag{2.2.1}$$

where matrix P includes statistically independent biological processes, and matrix $\boldsymbol{A}$ is a mixture matrix showing for each gene what biological processes contribute to the expression profile of the gene. In case of linear ICA,

$$D = f(AP) = AP \tag{2.2.2}$$

In order to find matrix $\boldsymbol{P}$, the linear ICA problem may me formulated as follows:

$$P \sim Y = WD \tag{2.2.3}$$

where we need to find a matrix $\boldsymbol{W}$ (called the unmixing matrix), so that rows of matrix $\boldsymbol{Y}$ are as statistically independent as possible. In this case, $\boldsymbol{Y}$ will be a close approximation for $\boldsymbol{P}$, up to permutation and scaling.

The process of finding the unmixing matrix can be performed by different algorithms, based on different metrics of statistical independence. For example, maximum likelihood estimation, a statistical approach for finding estimations of unknown parameters that result in the highest probability for observations [134], can

be applied. Another approach is to maximize negentropy (or equally minimize mutual information), given by the following equation:

$$J(Y) = H(Y_{gauss}) - H(Y), \text{ where } H(x) = -\int f(x)\log(f(x))dx \quad (2.2.4)$$

Maximum non-gaussianity can also be used as a measure of independence by using the Kurtosis metric:

$$kurt(Y) = E\{Y^4\} - 3(E\{Y^2\})^2 \quad (2.2.5)$$

where $E(Y^4)$ and $E(Y^2)$ are the 4th and 2nd moments of $\boldsymbol{Y}$ correspondingly.

As has been mentioned earlier, ICA has the advantage over PCA of not imposing a requirement of recovered signal orthogonality, and is therefore more favorable for recovering mixed signals. ICA also has been shown [135] to outperform PCA, k-means clustering and the Plaid model on combined yeast cell-cycle [120], yeast stress [92], *C. elegans* [136], and human normal tissue data [137]. Overall, ICA is a fast, robust algorithm that is very well suited for microarray analysis.

Although the statistical independence requirements of ICA is not as strict as orthogonality requirements of PCA, the assumptions about the independence of underlying processes may not be fully applicable in most microarray experiments. The method also does not take into account any error measures associated with microarray measurements and does not allow incorporation of prior knowledge into the analysis, leaving it prone to dimensionality problems from the large number of genes and minimal number of conditions generally under consideration.

2.3 Non-negative Matrix Factorization

First introduced for facial feature recognition by Lee and Seung [138], non-negative matrix factorization (NMF) was adopted for analysis of gene expression data. Various microarray analysis, including yeast mutants [139], classification of lung squamous cell carcinoma [140], analysis of leukemia [141] and toxicology datasets [142], have been analyzed using NMF since then.

NMF operates on preprocessed data from a set of expression array experiments. The data comprises estimates of mRNA transcript levels (single channel) or ratios (two channel) represented as a single matrix $\boldsymbol{D}$. Each row of $\boldsymbol{D}$ contains the mRNA estimates for each gene in all conditions (e.g., distinct tissues, experiments, timepoints), and each column corresponds to the estimates of mRNA levels for all genes in a single condition. For a dataset comprising I genes with expression measured in J conditions, the dimensionality of matrix $\boldsymbol{D}$ would be $I \times J$. The goal of the NMF simulation is to find a small number of metagenes (the number of metagenes provides a dimensionality estimate), each defined as a positive linear combination of I genes. The mRNA level estimates across conditions for each gene can be approximated then as a positive linear combination of these metagenes. Mathematically, this can be expressed as an approximate factorization of matrix $\boldsymbol{D}$ into a pair of matrixes $\boldsymbol{A}$ and $\boldsymbol{P}$ as in equation 2.3.1:

$$\mathbf{D} = \mathbf{M} + \varepsilon = \mathbf{A}\cdot\mathbf{P} + \varepsilon \tag{2.3.1}$$

The mock data, $\boldsymbol{M}$, is the approximation of $\boldsymbol{D}$, based on our estimates of $\boldsymbol{A}$ and $\boldsymbol{P}$. The matrix $\boldsymbol{\varepsilon}$ provides for the error in the measurements in $\boldsymbol{D}$. For K metagenes (i.e.,

K dimensions), matrix **A** is of size $I \times K$ with each of the K columns defining a metagene. The value of element A_{ik} indicates how strongly gene i is associated with metagene k. Matrix $\boldsymbol{P}$ is then of size $K \times J$, with each row representing the relative mRNA levels of a metagene across the conditions. The value of element P_{kj} given the strength of metagene k in condition j.

For NMF simulation, random matrices $\boldsymbol{A}$ and $\boldsymbol{P}$ are initialized according to some scheme. For instance, they could be populated from a uniform distribution U [0,1]. The two matrices are then iteratively updated using the following rules:

$$P_{\alpha\mu} \leftarrow P_{\alpha\mu} \frac{\sum_i A_{i\alpha} D_{i\mu}}{\sum_i A_{i\alpha} M_{i\mu}} \tag{2.3.2}$$

$$A_{\delta\alpha} \leftarrow A_{\delta\alpha} \frac{\sum_j D_{\delta\alpha} P_{\alpha j}}{\sum_j M_{\delta\alpha} P_{\alpha j}} \tag{2.3.3}$$

$$M_{ij} = \sum_k A_{ik} P_{kj} \tag{2.3.4}$$

which guarantees reaching a local maximum in Likelihood and minimizes

$$\left\| D - M \right\|^2 = \sum_{ij} (D_{ij} - M_{ij})^2 \tag{2.3.4}$$

In comparison to other factorization methods, NMF is capable of finding smaller, more localized patterns as well as global patterns [139], since it is doesn't require special properties of recovered metagenes. The only assumption is a non-negativity of the underlying signals that perfectly reasonable for additive nature of gene regulation. However, absence of such constraints has a tendency for recovering of signal-invariant metagenes that carry no or little information. This problem was addressed by Carmona-Saez *et al.* who developed modification to the method, non-

smooth NMF (nsNMF) to produce sparse representation of the metagenes and encoding vectors by making use of non-smoothness constraints [143]. Although very robust, NMF do not account for uncertainty information of the data, providing an issue of overfitting the data, just as PCA or ICA. Recent modification of the method, least-squared NMF (LS-NMF), introduced new updating rules for matrices recalculation by incorporating error measurements and replacing criteria of distance minimization with minimization of chi-square error [144] to take advantage of uncertainty information.

2.4 Bayesian Decomposition

The Bayesian Decomposition (BD) algorithm initially was applied to spectral imaging [145] and then was adapted to analysis of gene expression data [45]. It was successfully applied recently for expression pattern recognition in yeast deletion mutant gene expression data [146], yeast cell-cycle data [45], mouse tissue specific expression data [147], lung adenocarcinoma microarray data [148], and *Plasmodium falciparum* life cycle expression data [149].

The basic principal behind the decomposition performed by BD is to recover an amplitude matrix (***A***) and a pattern matrix (***P***), the product of which yields a model of the data matrix (***M***) that reproduces the data matrix (***D***) within the noise level (Eqn. 2.3.1), as shown in Figure 7. The initial matrix ***D*** represents measurements with errors for genes (rows) across different conditions or samples (columns). The recovered matrix ***P*** contains patterns of expression (rows) within the data and matrix ***A*** indicates the strength of the assignment of a gene to a pattern, thereby providing a

key feature to reflect the biological fact that one gene can be involved in multiple processes and have multiple patterns of expression.

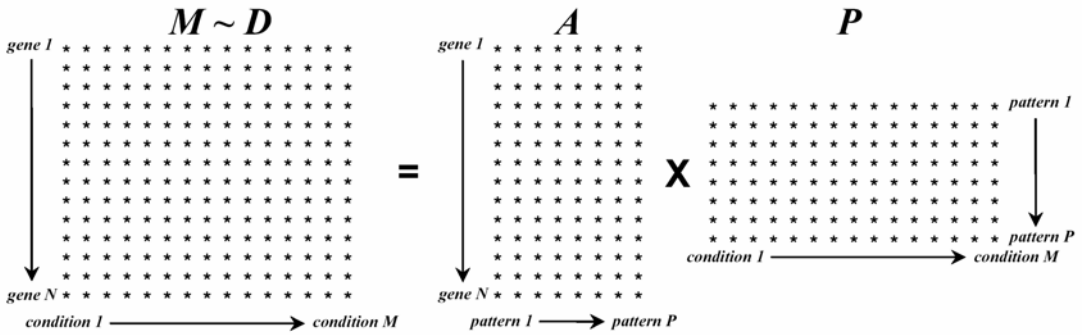

Figure 7. Decomposition performed by BD. Model matrix *M* is created by multiplication of matrices *A* and *P* recovered during the analysis to compare with initial the data matrix *D* **[45]**.

The algorithm creates amplitude and pattern matrices using a Markov Chain Monte Carlo (MCMC) process based on Bayesian statistics [150]. Bayes' equation is used to calculate conditional probability at each point of the Markov Chain, when changes in matrices ***A*** and ***P*** are created. That is, the probability of created matrices ***A*** and ***P*** to be a solution for the problem given the data ***D*** can be described by

$$p(A,P|D) \sim p(D|A,P)\, p(A,P) \qquad (2.4.2)$$

where p(A,P|D) is the probability of the current model (posterior), p(D|A,P) is the likelihood, and p(A,P) is the prior probability of the created matrices ***A*** and ***P*** to be a solution independent of the data ***D***. Each step of the MCMC process tests random changes made to matrices ***A*** and ***P*** according to the prior by determining the changes in the likelihood. Simulated annealing [151, 152] is used before sampling starts to

minimize the possibility of being trapped in a local maximum in the posterior distribution. That is done by modifying (2.4.2) to

$$p(A,P|D) = p(D|A,P)^T p(A,P) \quad (2.4.3)$$

where T is changed from 0 to 1, gradually increasing the influence of data on the posterior. After sampling begins, the MCMC process iterates for a given number of steps and returns the mean and standard deviation for each element of matrices ***A*** and ***P*** calculated from collected samples.

The process of MCMC with application of Bayes' formula can be implemented by creating an atomic domain with prior distributions and mapping to a domain of the matrices ***A*** and ***P***. This provides an ability to encode prior biological knowledge into the model. In the original work [45] prior knowledge encoded was the positivity of gene expression (i.e. the ratio of the relative amounts of mRNA between experiment and control is a positive value) and abundance of low gene expression signals compared to higher ones. The prior is composed by using atomic domain abstraction. The atomic domain is a set of atoms that have an amplitude (or value) and a position on an infinitely divisible line (actually 2^32 points). Changes that could be applied to the atomic domain include creating an atom of random amplitude (exponential, (1/q)e^(-z/q) with q being mean flux) and putting it to a random position (uniform); deleting an existing atom or moving an existing atom from one position to another. Thus, probability space formed by these atoms

represents positive additive distribution that also reflects prior information about microarray signal, giving more probability to atoms with lower amplitudes.

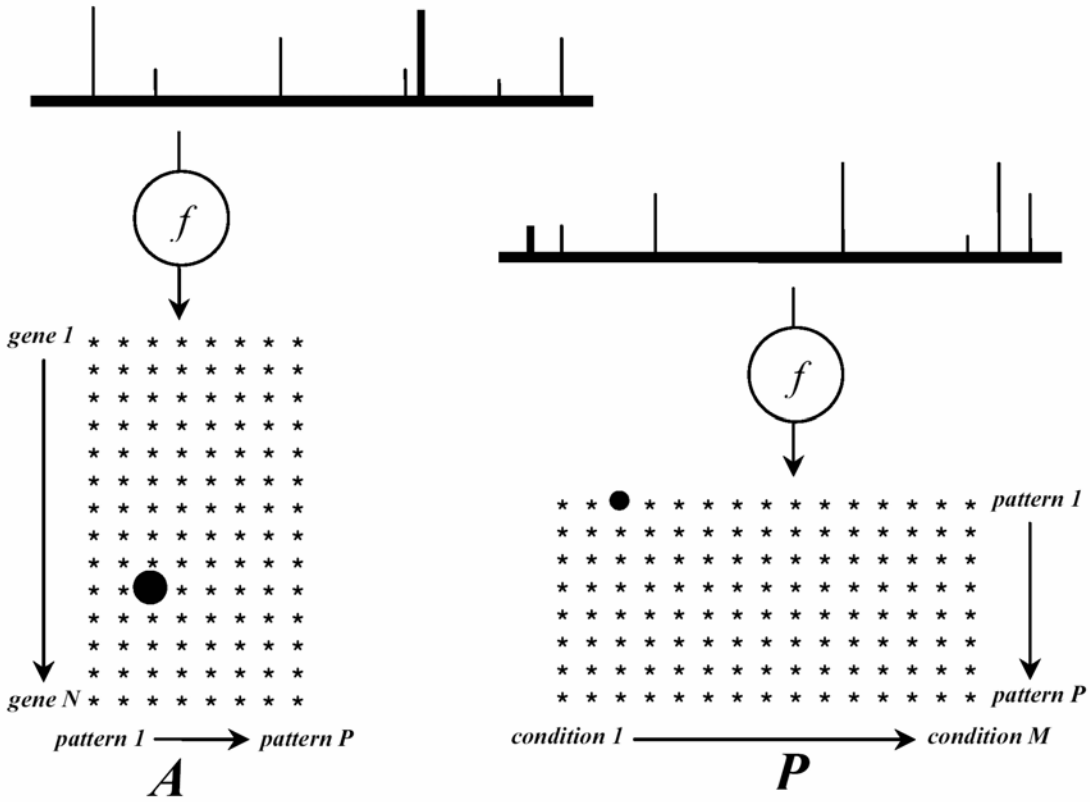

Figure 8. Creation of the prior by mapping atomic domains into the model. Mapping is done for each atom from atomic domain by convolution functions (*f*s), defined on atoms positions and amplitudes. Convolution functions shown simply map an atom to one element of a matrix, i.e. atoms denoted by thick lines would be mapped to the appropriate elements of matrix denoted by thick dot [45].

Atoms from the atomic domain are mapped to the model domain (matrices ***A*** and ***P***) by using a convolution functions (Figure 8). A convolution function maps the amplitude of each atom from the atomic domain into the values of one or more elements in ***A*** or ***P***. This mapping can be arbitrarily general. For example, a very simple convolution function is one that maps the amplitude of the atoms into a value

for one element of the target matrix [45] (Figure 9). Another convolution function may map one atom to a set of elements in matrix A that correspond to genes known to be co-expressed and therefore provide a mechanism for encoding additional prior knowledge into analysis.

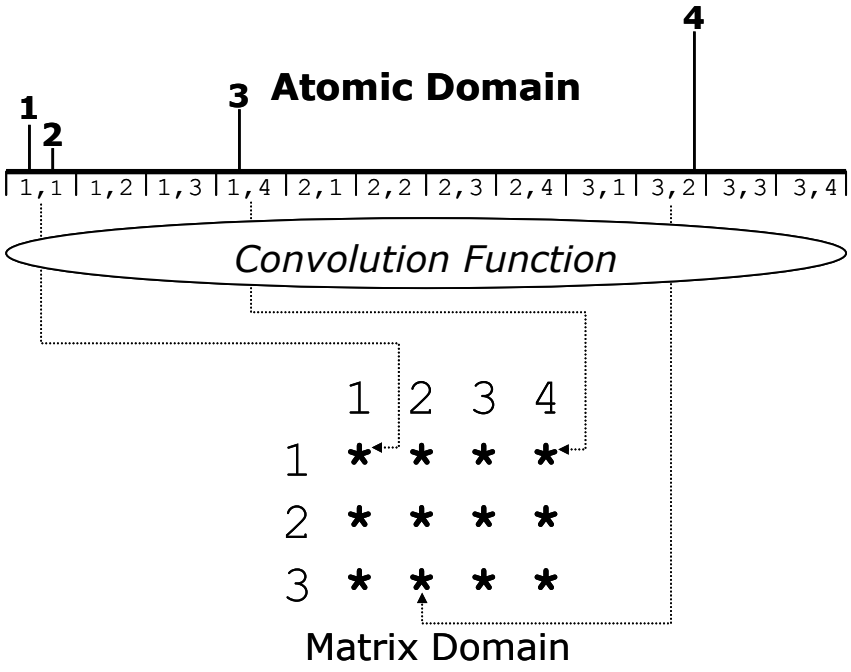

Figure 9. Mapping of an atom by a simple convolution function. Atomic domain is divided on a number of bins equal to number of matrix elements, with each bin correspond to the specific matrix element. Thus, an atom's amplitude is mapped to a matrix element that is determined by position of the atom. Atoms, 1 and 2 are mapped to the same element, while atoms 3 and 4 are mapped to different elements.

2.5 Summary of reviewed methods

All the reviewed advanced methods based on the model of matrix factorization allow assigning one gene to multiple groups of co-expression. In a process of matrix

factorization, it is possible to map recovered signals to expression patterns, which when mixed according to another matrix represent observed expression profiles for all genes from the initial data matrix. This way, PCA, ICA, NMF and BD are more favorable compared to clustering methods for analysis of microarray data, where biology dictates that genes will be regulated to function in more than one biological process.

It should be noted that assumptions for recovering of underlying signal can limit the applicability of the methods, e.g. orthogonality assumption of PCA or independance requirement of ICA may stop the methods to produce biologically meaningful patterns. While BD encodes information for expression patterns in a form of a prior, favoring patterns with a minimal structure, NMF does not have any such restriction except for non-negativity [143]. Absence of assumptions about underlying expression signals can also be a disadvatage when dealing with high noise in data, when outliers or missing points may lead the model to overfit the data.

Data overfitting is also inevitable when no noise model is introduced in the analysis. The original PCA, ICA and NMF methods do not have tools to handle variance measurements on microarray data, when modifications of PCA and NMF that target the issue were shown to improve performance of these methods.

The ability to include additional prior knowledge in the model is yet another desired property of a pattern recognition method for microarray analysis. This helps to address the issue of low the signal-to-noise ratio by including independant information in the model, allowing improving of the statistical power of results.

Bayesian Decomposition has an advantage over other methods, since it has mechanisms for such prior biological information inclusion in the form of the prior.

CHAPTER 3: ENHANCEMENTS TO BAYESIAN DECOMPOSITION

3.1 Modification of Bayesian Decomposition with Coregulation

Bayesian Decomposition uses a Markov Chain Monte Carlo process to create amplitude (***A***) and pattern (***P***) matrices that when multiplied together result in a model of the original data matrix (***D***) (Figure 7, section 2.4). While matrix ***D*** contains observed expression profiles, matrix ***P*** represents a suggested set of expression patterns (each row corresponds to a pattern) that relate to underlying biological processes, and matrix ***A*** indicates combination of which of these patterns yield the observed expression profile for each gene. Thus, the amplitude matrix ***A*** assigns each gene to patterns with varying strength, allowing one gene to belong to more than one pattern. A column of the amplitude matrix ***A*** can be interpreted as a group of genes that contain corresponding expression patterns from the pattern matrix ***P***, i.e. as a group of co-expressed genes. Prior information about such gene co-expression can be used in the process of creating the amplitude matrix ***A***, helping the algorithm to group together genes that have a high probability of being co-expressed.

As described in section 2.4, prior information can be encoded into the Bayesian Decomposition algorithm in the form of probability distributions for the amplitude and position of an atom or using a convolution function that maps atoms onto the matrix model (Figure 8). In order to encode the co-expression information we created a convolution function that maps the atom into a set of ***A*** matrix elements corresponding to the genes that are known to be coregulated (Figure 10). The process of defining target matrix elements is similar to the one shown in Figure 9, except each

bin is defined by a unique pair (co-expression group k, pattern p) and corresponds to multiple matrix elements. Thus, the position of each atom from the atomic domain falls into a bin (k, p), which indicates the convolution function for co-expression group k and pattern p, i.e. column p in matrix $\boldsymbol{A}$. The amplitude of the atom is mapped into the defined elements according to normalized weights for each gene. In order to reduce the negative consequences of a poor or incomplete prior in the model, the atomic domain is split onto two parts. For the first part a convolution function that maps an atom to a set of elements defined by coregulation information is used, and for the second part each atom is mapped to a single element of the matrix $\boldsymbol{A}$.

Normalization for the prior information is an issue that arises because of genes having different levels of expression and being involved in different sets of processes. This can be illustrated by Figure 11, where possible expression profiles of three genes known to be regulated in phase G1 are shown. The expression pattern that corresponds to only phase G1 is plotted as a dot line, and it is apparent that amplitude matrix values for these genes should be proportional to their expression levels in G1 for a decomposition to be correct. Therefore, a convolution function that uses co-expression information should spread an atom between elements of matrix $\boldsymbol{A}$ proportionally to expression levels of corresponding genes within that coregulation group.

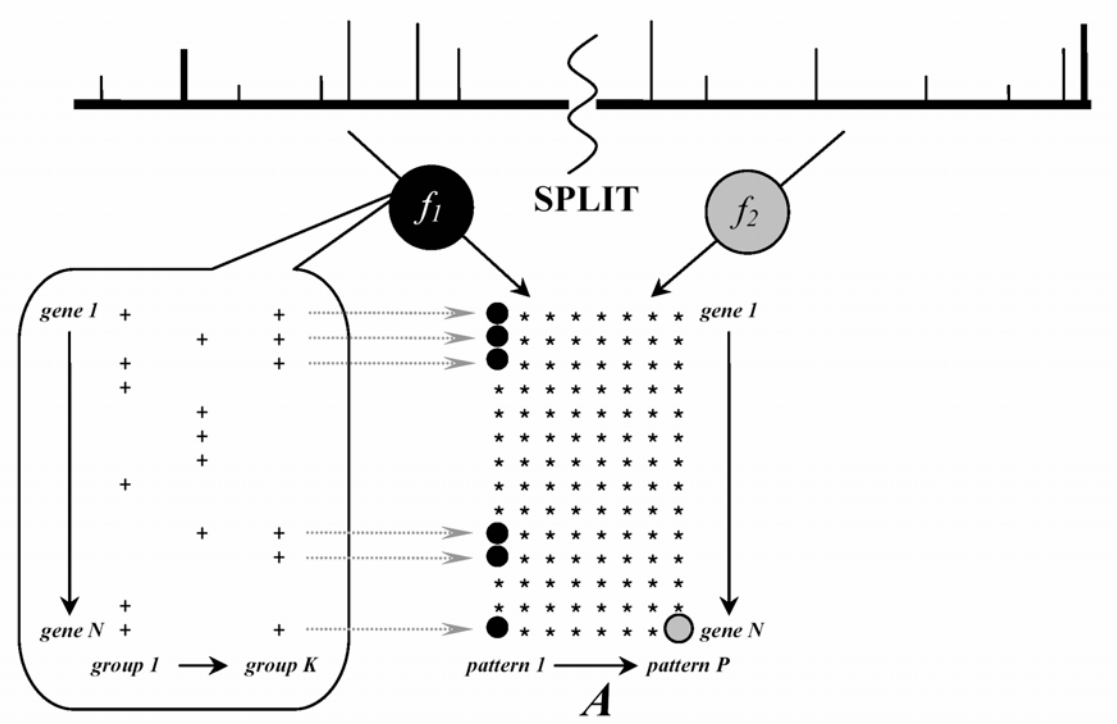

Figure 10. Splitting the atomic domain. The atomic domain is split onto two parts. The position of an atom (to the left or to the right of the split) in the atomic domain defines what convolution function will be used for its mapping. The convolution function that uses prior co-expression information (f_1) spreads the amplitude of the atom into elements in matrix ***A*** defined by the position of the atom (black dots in matrix A defined by position that resulted in using group K and pattern 1). The simple convolution function (f_2) is used to map an atom directly to appropriate element defined by the atom position (gray dot in matrix ***A***).

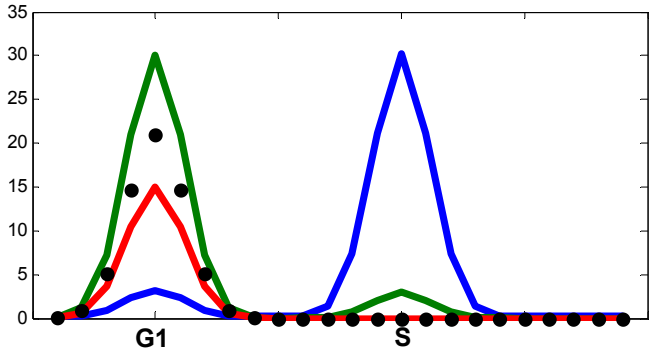

Figure 11. Expression profiles of genes coregulated in phase G1. Green, red and blue lines represent possible expression profiles of three genes. Dotted line shows expression pattern that corresponds to phase G1.

Given a number of coregulation groups *T*, we address the normalization issue by generating *T* overlapping subsets, with a subset *t* (t = 1..T) consisting only of data for genes from one group, and applying original Bayesian Decomposition to each of these subsets (Figure 12). The number of patterns posited into analysis for each subset is equal to *M* +1, where *M* is the total number of groups that contain any gene from the subset. This provides for a pattern for each coregulation set plus a pattern for routine metabolic function, which BD typically isolates in a separate pattern. The recovered amplitude matrices are used to determine the strongest pattern that explains the data subset: first, we normalize each row of matrix ***A*** (A_i) to the sum of 1:

$$A_i = \frac{A_i}{\sum_{j=1,P} A_{ij}}, \tag{3.1.1}$$

then, the strongest pattern corresponds to the column that gives the smallest variation coefficient (standard deviation over mean value) of normalized amplitudes. The strongest pattern $p^{strongest}$ is used to assign weights to each gene *k* of the group by calculating a dot product between the pattern and expression profile of the gene:

$$w_k = \sum_j d_{kj} p_j^{strongest} \tag{3.1.2}$$

The convolution function spreads the amplitude of an atom proportionally to the received weights for each gene from the co-expression group that corresponds to the atom.

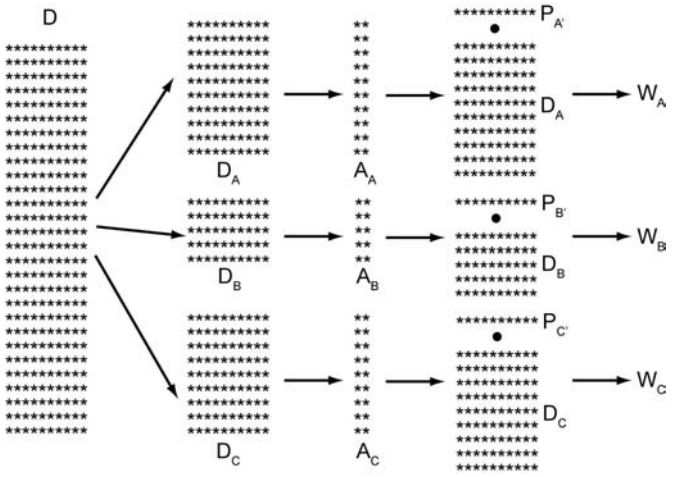

Figure 12. Calculating normalization weights. The figure shows example of the calculation process for three co-expression groups A, B and C. New subsets D_A, D_B and D_C are generated and BD is used for decomposition. After determining main pattern for each subset using recovered amplitude matrices A_A, A_B and A_C, dot products of main pattern and a subset are calculated to receives weight matrices W_A, W_B and W_C for each group of genes.

3.2 Testing Enhancements to Bayesian Decomposition

3.2.1 Introduction

We performed testing of the modified Bayesian Decomposition on three separate sets of data. First, we created a data set that simulates expression measurements of cell-cycle regulated genes. The simulation allowed us to control all the parameters, including underlying expression patterns, gene distribution between these patterns, and noise levels of modeled measurements. Next, we used a well studied yeast cell-cycle data set [90] (second data set) and yeast mutant (also known as Rosetta Compendium) data set [153] (third data set) to test enhancements done to BD on real microarray data. A wide knowledge base available for *S.cerevisiae* made it possible to acquire highly conservative regulation data about the genes under study, to compose prior information for modified BD, and to create a gold standard for

comparison of original and modified BD and other methods, such as hierarchical and k-means clustering.

In order to compare two algorithms between each other we used receiver operator characteristic (ROC) analysis [154] performed for groups of co-expressed genes received as a result of BD. Basically, a point of an ROC curve can be calculated by counting the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) assignments of all genes to groups and calculating the specificity and sensitivity values:

$$\begin{aligned} Sensitivity &= \frac{TP}{TP + FN} \\ Specificity &= \frac{TN}{TN + FP} \end{aligned} \tag{3.2.1}$$

Such groups used for calculating true and false positives and negatives are generally called a gold standard and have to be very reliable. To get other points of the ROC curve, analyzed algorithm should provide distinct grouping of genes by varying available parameters or thresholds. Then ROC curve is plotted as points of (1-specificity, sensitivity) and area under the curve (AUC) can be used as an efficacy measurement of the tested algorithm (Figure 13).

ROC analysis was performed for BD by increasing the stringency of assignment of a gene to a pattern. Essentially, each gene has a mean value of its strength within a pattern and an uncertainty on that assignment from amplitude matrix $\boldsymbol{A}$ based on the MCMC sampling. By increasing the number of standard deviations away from zero required to assign a gene to a group, multiple estimates of the assignment of the genes to the patterns were made, allowing the ROC curve to be constructed.

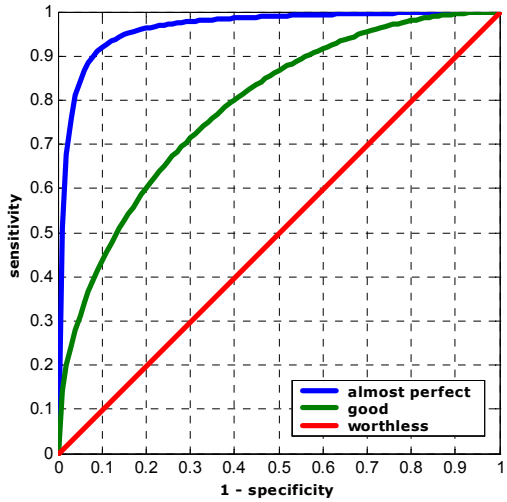

Figure 13. Examples of ROC curves. Blue, green and red lines represent three algorithms. Plotted as 1-specificity against sensitivity, blue curve correspond to the best of the three methods, green is less accurate than blue, and red is an example of algorithm that generates results randomly.

3.2.2 Testing on simulated data

First, validation of our modifications to Bayesian Decomposition was performed using a simulated gene expression data set modeling yeast cell-cycle. We created amplitude (***A***) and pattern (***P***) matrices (Figure 14) and multiplied them together to generate an 'ideal' data matrix. The pattern matrix included 5 overlapping patterns imitating expression profiles of 4 cell-cycle phases (G1, S, G2, M) and a metabolic oscillator taken through 2 full cycles (48 sample points) [45, 90]. Amplitude for cell-cycle imitating patterns had the value of 3 and metabolic oscillator alternated expressions of 0.15 and 0.1. The amplitude matrix was created to simulate expression profiles for genes whose expression is regulated at multiple phases of the cell cycle: 288 genes were randomly assigned with a combination of patterns, so that 126 genes were regulated in phase G1, 122 genes in phase S, 116 genes in phase G2,

115 genes in phase M and 64 genes had an expression pattern of a metabolic oscillator. Overall, 72 genes contained only one, 181 genes contained two, 31 contained three and 4 contained four expression patterns.

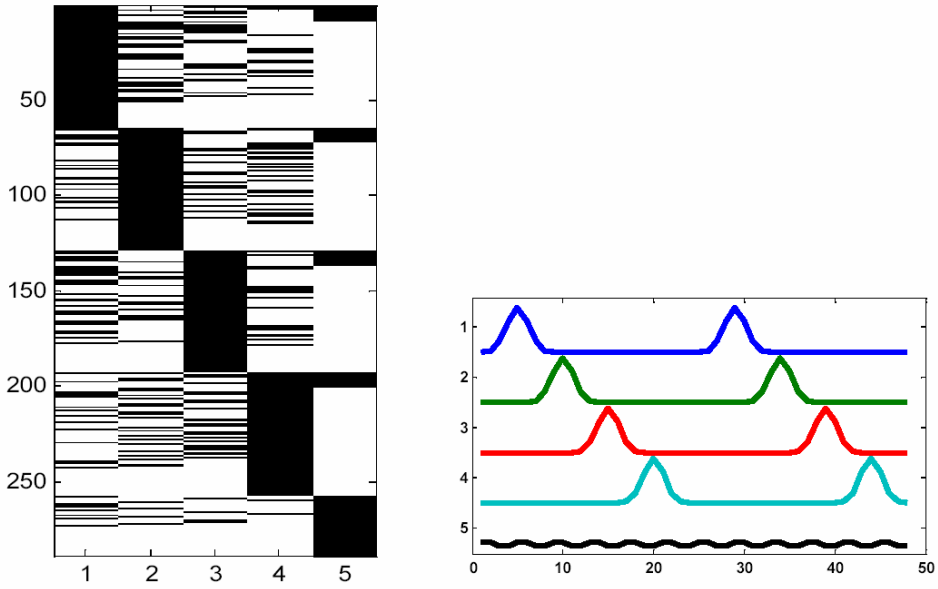

Figure 14. Simulated ***A*** and ***P*** matrices. Matrix ***A*** consists of 288 genes with expression linked to 5 patterns. Black stripes show if the gene shows expression related to the pattern. Matrix ***P*** consists of 5 patterns, simulating 4 cell-cycle phases and a metabolic oscillator.

To reflect noisy signals the data matrix was distorted, including different levels of additive and multiplicative noise as expected for microarrays [155, 156], i.e.

$$D = N(0,\sigma_a) + e^{N(0,\sigma_b)} \cdot (A \cdot P) \qquad (3.1.4)$$

where ***A*** and ***P*** are simulated amplitude and pattern matrices, and σ_a and σ_b are additive and multiplicative levels of noise respectively. Data matrices with 154 different noise levels were created, varying additive noise levels from 0 to 6.5 with step 0.5 and multiplicative noise levels from 0 to 3 with step 0.3 with the data matrix

without noise having maximum amplitude 3.15 and mean amplitude 0.65. While peak expression levels in the ideal matrix D had a value of 3.15, variation of the noise levels provided noise coverage (noise levels of more than 100% of signal σ_a=6.5 and σ_b=3) that is far beyond error levels of real microarray data (10-30% of signal). For each noise level 4 replicates of matrix ***D*** were created representing an experiment with four replicates and mean and standard deviation of mean were used for the simulation.

The 154 data matrices with different levels of noise were processed with original and modified versions of BD. Each data matrix was analyzed using four different random seeds (different starting points for Markov Chain Monte Carlo) positing 5 patterns in the model, which corresponds to:

$$D_{ij} \sim M_{ij} = \sum_5 A_{ik} P_{kj}, i = 1..288, j = 1..48 \quad (3.1.5)$$

Prior knowledge of gene co-expression provided to the modified BD algorithm was composed by taking information from the simulated matrix A with different coverage – 1 (all information – 5 groups with 126, 122, 116, 115 and 64 genes), 0.9 (5 groups with 113, 113, 103, 104 and 50 genes), 0.75 (5 groups with 100, 92, 87, 80 and 54 genes), 0.5 (5 groups with 64, 62, 55, 59 and 39 genes) and 0.25 (5 groups with 30, 31, 23, 29 and 16 genes).

Results were compared for each level of noise and for each level of prior information included between the original and modified BD. The χ^2 fit between simulated and calculated amplitude matrices and the value of the area under the ROC curve for grouping the genes determined from the amplitude matrix were used for the

comparison. Figure 15 shows results in a form of heat maps where orange-red blocks indicate an advantage of using the modified version of BD at the given noise level and blue blocks indicate a disadvantage (green indicates no significant difference). The results for simulated data demonstrate that increasing the amount of available prior information about gene coregulation in the analysis allows the modified BD more accurately find patterns in noisy data. Although there are rare points of noise levels where original BD performs better, typical modern arrays have noise levels in the lower left quadrants and including prior information always showed better results at these levels.

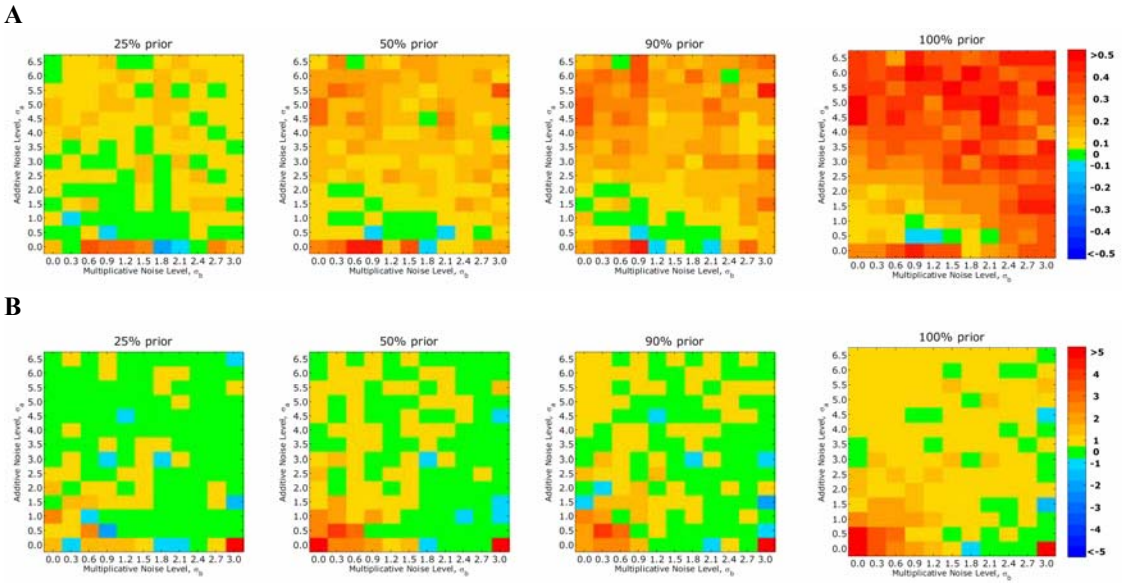

Figure 15. Comparison of original and modified Bayesian Decomposition based on simulated data.
Red-yellow bars identify levels of noise where modified BD showed better results, and magenta-blue bars identyfy where original. Green bars show insignificant difference. **A**. Figure shows log2 ratios between chi-squared distances for original and modified BD with different amount of prior information included (25%, 50%, 90%, and 100%). Chi-squared is calculated between ideal and recovered amplitude matrices. **B**. Figure shows log2 ratios of areas under ROC curve for original and modified BD with different amount of prior information included (25%, 50%, 90%, and 100%). Areas under ROC are calculated using information of ideal amplitude matrix as gold standard for recovery results.

We also analysed number of atoms required for the algorithm to fit the data at different noise levels. At each noise level, average number of atoms used by BD for the amplitude (N(A)) and pattern (N(P)) matrices were received from MCMC sampling for both original BD and modified BD when all prior information was used. Figure 16 presents histograms that show distribution of differences between number of amplitude matrix atoms (*A* atoms, on the left) and between number of pattern matrix atoms (*P* atoms, on the right) used by original and modified BD to fit each of 154 data matrices. While histogram for pattern matrix show no difference of *P* atoms used by both versions of BD (average difference value of 0), average number of *A* atoms used by modified BD (643) is less than average number of *A* atoms used by original BD (753) yeilding the average difference value of 110. These results indicate that the enhancements made to the BD algorithm work as intended, requiring to create less amount of atoms, in other words less amount of information, to find the amplitude matrix.

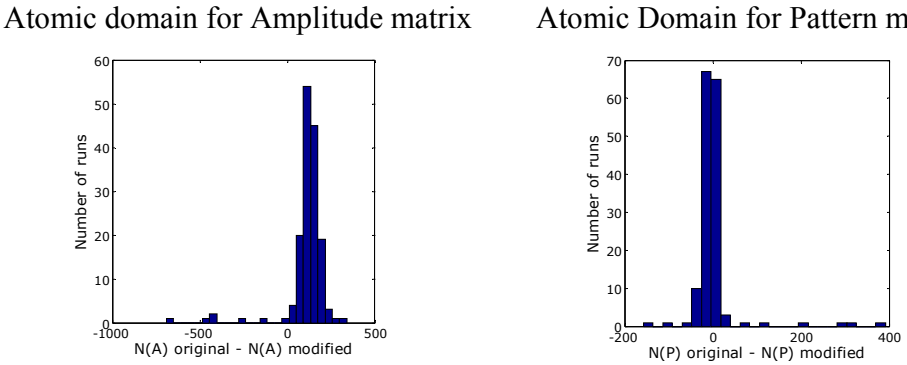

Figure 16. Histograms of atoms number differences between original and modified BD. Histograms are built based on 154 values, one for each level of noise. Atoms number difference for atomic domain that corresponds to the amplitude matrix is on the left. Atoms number difference for atomic domain that corresponds to the pattern matrix is on the right.

3.2.3 Testing on biological data: yeast cell-cycle dataset

Validation of the modified Bayesian Decomposition on simulated data showed that incorporation of prior coregulation information into the analysis can significantly increase the ability of the algorithm to recover the correct groups of co-expressed genes. While simulated data provides unique control over all testing conditions, including most importantly knowledge of the correct results, a biological example needs to be analyzed to confirm the power of the algorithm. The problem of the lack of a gold standard measurement can be partially overcome by choosing a biologically well-characterized system. Budding yeast (*S.cerevisiae*) is a well-studied model organism with standard data sets analyzed by various algorithms, providing an additional opportunity to check the efficacy of BD against other methods.

The cdc28 temperature-sensitive mutant yeast cell cycle data [90, 120] comprises measurements of gene expression levels over 160 minutes covering two complete cell cycles. Patterns of expression found within the data can be mapped to cell-cycle phases and distribution of genes among these patterns makes it possible to group genes into co-expression sets. The yeast cell cycle data set was widely used to test different algorithms including singular value decomposition [119], independent component analysis [129], cooperative vector quantizer [157] and shrinkage-based cluster analysis [94]. The latter has a receiver operating characteristic (ROC) curve [158] built on the results of clustering analysis that is compared with ROC curve built on the results from BD.

Yeast cell cycle data comprises 788 cell-cycle regulated genes with expression measured across 17 time points. Missing points were replaced with value of 1.0 with uncertainty of 9079 - the maximum value of the dataset, so BD wouldn't be constraint by such points. The prior knowledge about genes coregulation were taken from literature by searching evidences of genes regulation by one transcription factor that wasn't based on microarray data. It resulted in 11 groups contained from 5 to 17 genes in each, 67 genes in total and 18 from them belong to more that one group (see Table 1).

Table 1. Coregulation data [94] used for analysis of yeast cell-cycle data

Group	Regulator	Target genes
1	Mot3	*Hxt4, Suc2, Cyc1, Sst2, Hxt2*
2	Ndt80	*Clb1, Clb6, Clb4, Sps4, Clb5*
3	Ste12	*Tec1, Fus1, Far1, Cln1, Mfa2*
4	Swi5	*Pcl2, Pcl9, Ash1, Sic1, Egt2*
5	Cbf1	*Met10, Met28, Met3, Met17, Met25, Met16*
6	Fkh1	*Swi5, Alk1, YIL158W, Bud4, YPL141C, Clb2*
7	Fkh2	*Swi5, YIL158W, Bud4, Ace2, YLR190W, YPL141C, Kip2, Clb2*
8	Swi6	*Ho, Rnr1, Swi4, Cdc6, Cln1, Cdc21, Cln2, Clb5*
9	Mcm1	*Cln3, Swi5, Mfa1, Swi4, Ste2, Far1, Cdc6, Ace2, Cdc46, Clb2*
10	Swi4	*Pcl2, Ho, Mnn1, Och1, Cis3, Cwp1, Gls1, Cln1, Pcl1, Srl1, Svs1, Cln2, Kre6*
11	Rlm1	*Pst1, Sed1, Crh1, Mpk1, Sec28, YIL117C, Cis3, Pir2, Cwp1, Pir3, Pir1, YLR194C, Fks1, Dfg5, YMR295C, YNL058C, Ygp1*

Number of patterns posited into the analysis was from 4 to 7 and reflected the nature of the data to contain just cell-cycle regulated genes. We analyzed the data set with both the original and modified BD and results were compared to determine the effect of prior knowledge. We used two different gold standards for the ROC analysis. First gold standard consisted of prior knowledge groups from Table 1 used

by modified BD for the analysis. Second gold standard was based on the known molecular biology of gene coregulation independent of microarray studies and comprised 9 groups with 43 genes total used by Cherepinsky *et al.* [94] to build the ROC curve for hierarchical clustering algorithm available to compare with both BD methods.

In Figure 17 we present results of the application of modified BD to the cell cycle data using ROC analysis. We compared the results using the original BD (circles), modified BD (squares), and shrinkage-based hierarchical clustering (triangles) performed previously [94]. On the left, the ROC curve is build using gold standard from Table 1 and shows that prior coregulation information was successfully used by modified BD to provide the results of better accuracy than original BD. On the right, Cherepinsky *et al*. groups were used as gold standard and modified BD obtains an area under the curve of 0.82, compared with 0.83 for original BD and 0.56 for the best hierarchical clustering method. The lack of improvement from use of coregulation information reflects the limited nature of such data for these genes at the present time. In this case, we have prior data on only 67 of 788 genes, which was not adequate to improve inference over BD. However, we include the results to show the value of both original and modified BD, due to assignment of genes to multiple groups.

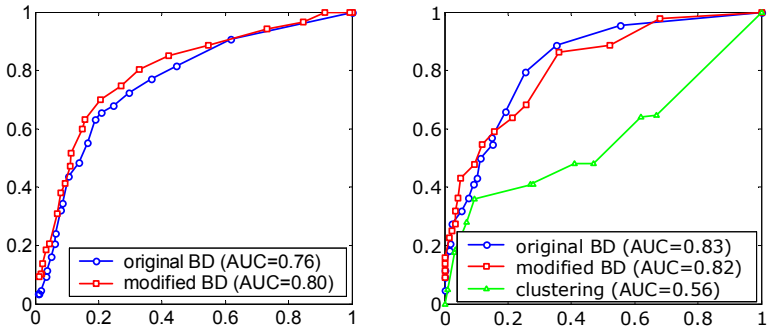

Figure 17. Results of yeast cell-cycle analysis. Left figure shows ROC curves for original and modified BD based on golden standard from Table 1. Right figure shows comparison of ROC curves for BD with hierarchical clustering based on groups from Cherepinsky et al. [94].

3.2.4 Testing on biological data: Rosetta compendium dataset

The Rosetta Compendium data set [153] contains microarray profiling of 300 diverse mutants and chemical treatments in *S.cerevisiae* and of 63 wild type control samples grown in rich media. Because *S.cerevisiae* has several MAPK signaling processes, expression patterns recovered from this data may be mapped to signaling pathways. Specific mutants with a knocked-out gene in a key pathway component offer the opportunity to validate the transcriptional patterns related to pathways.

The data set was analyzed with both the original and modified BD algorithms and an ROC analysis was performed to compare efficacy of both algorithm versions. The gold standard set used in the yeast cell cycle data set could not be used here as the specific genes show little variation in the Rosetta data set, which is not unexpected as all cultures were actively growing and unsynchronized, and therefore they showed active but average levels of expression of the cell cycle genes.

Coregulation information included as prior knowledge in the modified version of BD was obtained similarly to yeast cell-cycle data from biological literature based on reported transcription factors and their target genes, determined with methods other than microarrays (Table 2). The same data was used as the gold standard for the ROC analysis. We ran with both the original and modified BD positing from 10 to 20 patterns into the analysis.

Table 2. Coregulation data used for analysis of Rosetta compendium data

Group	Regulator	Target genes
1	Zap1	*Adh4, Oye3, Zrc1, Zrt1, Zrt2*
2	Ndt80	*Clb1, Clb6, Dit1, Sps1, Sps4*
3	Mcm1	*Clb2, Far1, Mfa1, Mfa2, Ste2, Ste6*
4	Gcn4	*Atr1, His3, His4, His7, Ilv2, Rad16*
5	Dal80	*Dal3, Gap1, Gdh1, Put1, Put4, Uga4*
6	Rtg1	*Aco1, Cit2, Idh1, Idh2, Pdr3, Pdr5, Pox1*
7	Pdr1	*Pdr10, Pdr15, Pdr3, Pdr5, Ste6, YAL061W, YLR346C*
8	Met4	*Ecm17, Met14, Met16, Met17, Met17, Met3, Met6*
9	Ume6	*Dit1, Ime1, Ime2, Ino1, Opi3, Sip4, Spo13, Sps2*
10	Ste12	*Far1, Fus1, Mfa1, Mfa2, Muc1, Pgu1, Ste2, Tec1*
11	Mot3	*Dan1, Hxt2, Hxt4, Leu2, Sst2, Suc2, Tir1, Tir4*
12	Gln3	*Dal3, Gap1, Gdh1, Gdh2, Put1, Put4, Uga4, Ura3*
13	Cbf1	*Gal2, Met10, Met14, Met16, Met17, Met17, Met3, Sam2*
14	Mig1	*Dsf1, Emi2, Hxk1, Hxt13, Hxt2, Hxt4, Reg2, Suc2, YFL054C, YLR042C*
15	Rlm1	*Crh1, Ctt1, Cwp1, Fit2, Pir3, Prm5, Pst1, Sed1, Slt2, Sps100, Ygp1, YLR194C*
16	Msn4	*Adh2, Ahp1, Glk1, Gph1, Gre2, Gsy1, Hsp104, Hsp26, Hsp30, Hxk1, Ime1, Msc1, Pgm2, Rnr3, Rtn2, Sol4, Spi1, Suc2, YNL194C*
17	Msn2	*Ahp1, Ctt1, Glk1, Gph1, Gre2, Gsy1, Hsp104, Hsp26, Hsp30, Hxk1, Ime1, Msc1, Pgm2, Rnr3, Rtn2, Sol4, Spi1, Suc2, YNL194C*

Figure 18 demonstrates results of the analysis of Rosetta compendium data. Results from using modified BD (squares) were compared to K-means clustering (triangles) and the original BD analysis (circles) of the same data. Here, both the

coregulation information and the gold standard gene lists were the same, so that the results demonstrate that the algorithm correctly used information about transcription factor regulation and that such coregulation is reflected in the data. All techniques performed equally well at high specificity, however as sensitivity increased, modified BD was superior in terms of reduced false positives due to the inclusion of prior information on expected coregulation.

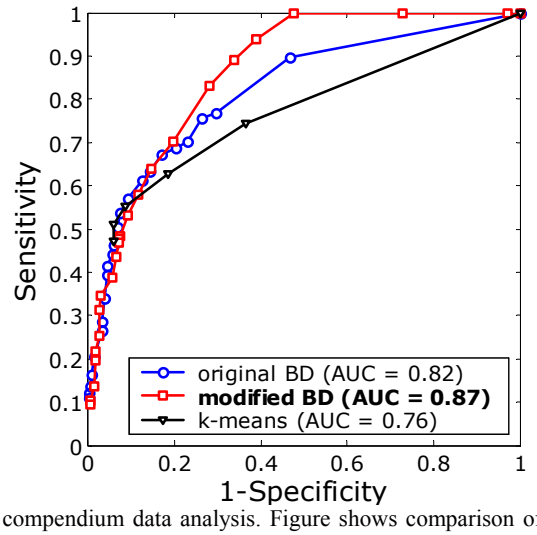

Figure 18. Results of Rosetta compendium data analysis. Figure shows comparison of ROC curves for original and modified BD along with k-means clustering.

3.3 Summary

Testing of enhancements done to Bayesian Decomposition on simulated data showed the advantage of using coregulation information, as it increases as levels of prior information increase. The observed behavior indicated improvement of statistical inference compared to the original BD algorithm. Reduction of A atoms

required for the algorithm to find correct solutions also indicated the success of development of modified Bayesian Decompostion. BD enforces an additional constraint on the prior and therefore there is a preference for minimization of structure, which takes the form of a tendency to use fewer atoms. This Occam's razor argument leads to relatively sparse matrices.

Testing performed with real biological data revealed that even with the small amount of prior coregulation information, modified BD performs equally well or better than original BD. Although yeast cell-cycle dataset analysis showed no difference in efficacy between both versions of the algorithm when compared to golden standard that was not included as prior knowledge, results check based on the data included as prior coregulation information demonstrated advantage of using such information in order to improve statistical power. Rosetta compendium dataset results interpretation showed that both algorithms performed equally well at high specificity, while as sensitivity increased, modified BD was superior in terms of reduced false positives due to the inclusion of prior information on expected coregulation.

This work demonstrated the value of inclusion of prior knowledge of transcriptional regulation in the analysis of microarray data and also the present limits on that knowledge. While the simulations showed a clear advantage in using this knowledge, the analysis of yeast data indicates the present lack of coregulation information available. Nevertheless, the superiority of the modified BD approach is clear. Our knowledge of transcriptional regulation is rapidly increasing, and we expect improved statistical power with modified BD over the next few years. This

power will be critical to improved inference of biological process activity, especially with heterogeneous and limited samples typical in clinical settings.

CHAPTER 4: AUTOMATED SEQUENCE ANNOTATION PIPELINE

4.1 Automated Sequence Annotation Pipeline concept

In order to maximize discovery, a vast amount of biological information about genes are available to guide microarray data mining, starting from identifying probes on microarray slides and ending with interpretation of results received by a data processing algorithm (Figure 4, section 1.6). While general annotations frequently can be done through various systems that provide web access to databases of interest, specific requirements usually demand using output of one query as an input for another, and often a researcher wishes to use different internet sources for the task. The annotation process therefore becomes time consuming, and the user is expected to have an expertise in different queried systems. The manual processing is also complicated if annotation is performed for a list of entries and the target sources do not support batch queries, especially when the list needs regular updating due to constantly changing information from annotation sources. Finally, an investigator needs to post-process received annotation information to organize it in a format required for data mining or result interpretation software.

In order to overcome most of the issues of manual annotation, we created the Automated Sequence Annotation Pipeline (ASAP) system [159] to automate acquiring of annotations. We use it as a part of a data mining process that supplies microarray data analysis at each step. The ASAP system serves as a mediator between the user and various data sources, using results from one query as input parameters for another to receive annotations for the user's data (Figure 19). Data

sources available for ASAP to query can be of various types and located in different sites. For example, remote dababases with web access by http or ftp protocols can be a valuable source of annotation information. Also, local programs can be used by ASAP to run them and receive results of calculations, which can form part of the information required by the user. Finally, local database can be used to store information once acquired from other sources. It allows caching results to provide quicker and more reliable access to them without querying remote databases that can be slow and even unavailable at the moment of request.

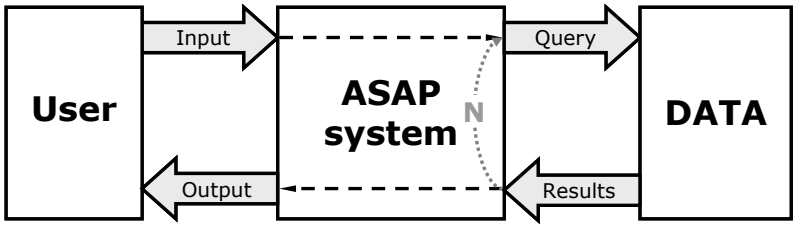

Figure 19. Automated Sequence Annotation Pipeline system. User passes input data and ASAP system performs a series of queries, using user's input and results from other queries. Output is formed from the results of queries and passed back to user.

ASAP performs annotation of user input data by executing pre-designed annotation plans that represent a script of directives, which include specifics of what data sources to query, what format the target source accepts query input, and how to extract information (results) from the received data. Also, annotation plans contain descriptions of formats available for user output. Thus, ASAP represents an environment where annotation plans are stored and run. In addition, the system generates reports for the administrator about possible changes in query input formats

or availability of remote sources. This is especially important for the system, since it assumes fixed standards that can change without notice.

Another feature in the system is the allowance for multiple users that can work on different operating systems. In addition, there is also a need to have up-to-date information for querying. Due to these demands, ASAP was designed as a web-application with a central linked local database, and it requires a researcher only to have an internet browser to access the pipeline. Two distinct web interfaces were created to separate functions available for regular users and administrators of the system. The user part provides an interface to basic functions of ASAP, such as submitting a job by querying an annotation plan of interest with parameters, checking status of the jobs and downloading results of annotations. The administrator part consists of various maintenance functions, including user management, basic system parameters configuration, detailed job information handling and, most importantly, new annotation plan installation and editing functions. A complete list of scenarios (also called use cases) available for user and administrator roles is shown in Figure 20, and descriptions of these use cases are provided in Table 3.

4.2 Automated Sequence Annotation Pipeline implementation

ASAP is implemented as a client-server web application using the Perl language as shown in Figure 21. Users access the system with the web interface through web browsers that support cascade style sheets, e.g. Internet Explorer, Safari, Opera or Firefox. The system can be installed on Linux or Windows platforms under the Apache server that supports execution of Perl scripts. On the back end ASAP uses a

relational database (mySQL) to store information about system users, performed tasks, status of jobs, and other user and administrator related information. A local database is also used to store annotation information acquired earlier. Various supported protocols of access allow ASAP core functions to access remote and local data sources both local and remote.

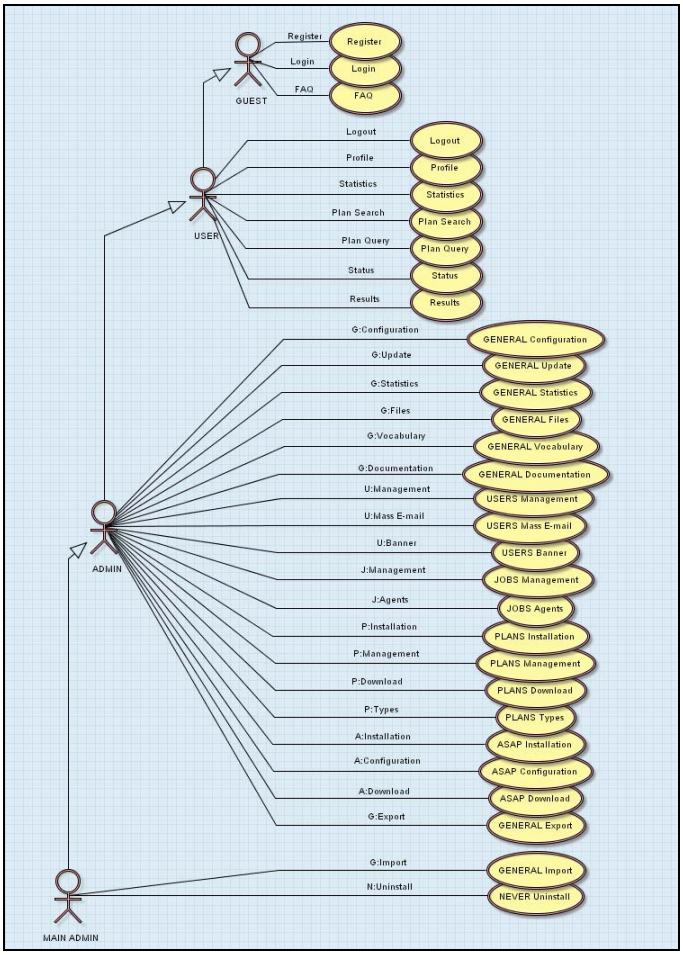

Figure 20. Use case diagram for ASAP system. Use cases available for administrators (MAIN ADMIN, ADMIN) and regular users (USER, GUEST) are shown and descriptions are provided in table.

Table 3. Description of ASAP use cases

Use Case	Description
Register	This use case allows GUEST to register in the system by setting username, password and e-mail.
Login	If GUEST is registered in the system the use case allows for GUEST to log in into the system and become a USER.
FAQ	This use case allows to see frequently asked questions page.
Logout	The use case allows USER to log out from the system. By this the USER becomes GUEST.
Profile	This use case allows USER to modify his/her profile information.
Statistics	This use case allows USER to see general statistics of his/her activity in the system, i.e. the amount of queried done to the system plans.
Plan Search	This use case allows USER to search for necessary plan by specifying search parameters.
Plan Query	This use case allows USER to query a plan to receive annotations.
Status	This use case allows to see the status of query submitted by the USER earlier.
Results	This use case allows to retrieve results of the query (if any) received for the USER's plan query.
GENERAL Configuration	This use case allows to modify various systems parameters.
GENERAL Update	This use case allows to update the system from installation package or patch.
GENERAL Export	This use case allows to create export file for the system that contains all plans, structure and data of www and asap tables (for agent_created tables it is only structure). The file is to be used in other system to make them absolutely the same as this one.
GENERAL Import	This use case allows MAIN ADMIN to update the system internal data to make it completely equivalent to the system the file to import is received from.
GENERAL Files	This use case allows to browse through files of the system and update/delete them.
GENERAL Statistics	This use case allows to see basic statistics for the work of the system for specified period of time.
GENERAL Vocabulary	This use case allows to view/add/modify terms of the system's vocabulary.
GENERAL Documentation	This use case allows to download the latest ASAP documentation.
USERS Management	This use case allows to search for a user by name and modify the found user's profile.
USERS Mass E-mail	This use case allows to send mass e-mails to groups of users.
USERS Banner	This use case allows to set text for the banner to be shown in each interface page of the system for users to see.
JOBS Management	This use case allows to search started queries by their job identificators and see the status of that jobs
JOBS Agents	This use case allows to to run/stop agents launcher - the program that delivers agent plans for execution with some periodicity automatically. Also, allows to change original parameters of such periodicity for agents and activate or deactivate such agents, or see next scheduled time for their execution.
PLANS Installation	This use case allows to install new plans into the system.
PLANS Management	This use case allows to search for a plan by various criteria and modify the system information about the plans.
PLANS Download	This use case allows to download the system plans to the user's desktop.
PLANS Types	This use case allows to add/modify/delete plans types.
ASAP Installation	This use case allows to install all copied (manually into the file system) plans into the system.
ASAP Configuration	This use case allows to change ASAP core parameters.
ASAP Download	This use case allows to download the ASAP core modules to users desktop.
NEVER Uninstall	This use case allows for MAIN ADMIN to uninstall the system completely.

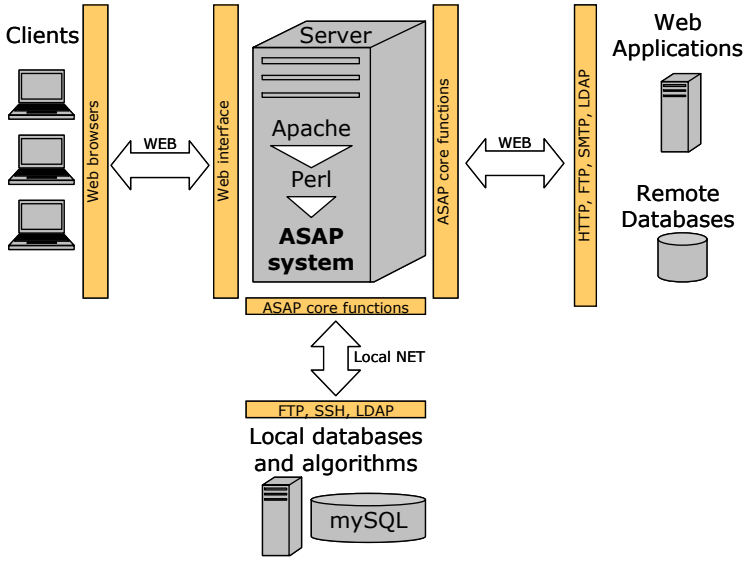

Figure 21. ASAP deployment diagram. Implementation of ASAP allows to install it as a stand-alone server under Apache web server software with Perl language support. Clients access ASAP though web browsers that interact with the system web interface. ASAP uses its core functions to access though HTTP, FTP, SMTP, LDAP, SSH and other protocols to remote web applications, databases and local algorithms and databases.

Figure 22 shows schematically a data flow for core ASAP functions that allow a user to submit a job to ASAP by querying an annotation plan, to check status of the job and to download results of annotations. The system comprises a set of modules that handle the generation of queries, parsing of results, formatting final reports, and various managerial functions. The user interacts with the system either through a web interface or through a configuration file in XML format. The inputs include a gene list (typically accession numbers) and a reference to a pre-designed plan for visiting multiple web sites. Plans can be easily created by an administrator as needed. The JOB MANAGER is a module that uses the database (DB) to generate the plan and then passes information to the plan. The Annotation Plan is a Perl script that can also

use other Annotation Plans to query local databases and external web resources, passing the results of one query to the input of another as necessary. ASAP stores received results in output files and information about these files in local database. The JOB MANAGER then uses this information to create download links and e-mails this information to the user. In case of any external source disruption, an error is generated and sent back to the user, while detailed error information is sent to the Administrator. The Administrator is then able to reply quickly to fix the query, if the error has occurred because of a change in web formats.

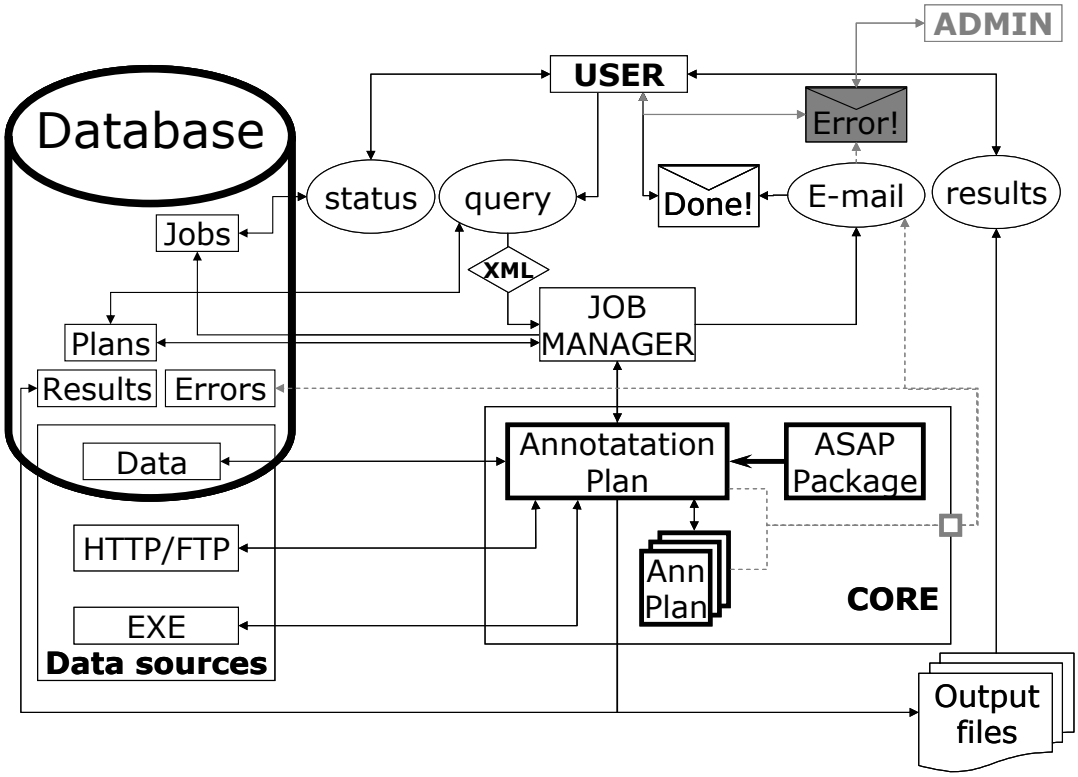

Figure 22. Data flow schema for ASAP core functions. JOB MANAGER handles USER's query request and passes input information to Annotation Plan that uses ASAP Package and possibly other Annotation Plans to query local Database, external data sources though HTTP or FTP protocol, or local executable algorithms to receive results of annotation. USER can check information about submitted job status and download results when output files are formed and available. E-mails are used to report results or possible errors to USER and ADMIN.

4.3 Automated Sequence Annotation Pipeline annotations

To support data mining process we created a set of annotation plans for ASAP that automatically acquire data required for each step of data analysis. These annotations include Affymetrix and Agilent probes identifications, UniGene [160] cluster names, descriptions and gene symbols for GenBank accession numbers, gene ontology terms and transcription factors for genes. In order to provide results with these annotation plans, ASAP uses various external data sources, including Affymetrix and Agilent updated slide information, the UniGene database, the Swiss-Prot protein knowledgebase [112], the Gene Ontology consortium [115] database and the professional version of TRANSFAC [161].

First, we implemented the UniGene annotation plan for GenBank accession numbers that retrieves corresponding information about UniGene cluster ID, gene name and description for provided GenBank sequence ID. Gene ontology information, such as what molecular function a gene performs, what biological process it is involved in and what cellular component it resides in, is also provided for the sequence. Figure 23 shows schematically directives for ASAP for the annotation plan. The plan is particularly useful for filtering and interpretation of results during microarray data analysis.

ASAP uses regularly updated annotation files from Agilent [162] and Affymetrix [163] to identify probes on a slide (assigns GenBank accession numbers) and uses UniGene plan to provide UniGene cluster ID, gene name and description along with gene ontology for the probe. The information received with this plan can

be uses at the first step of microarray data mining process before data pre-processing, for example, to filter out probes that do not represent know genes.

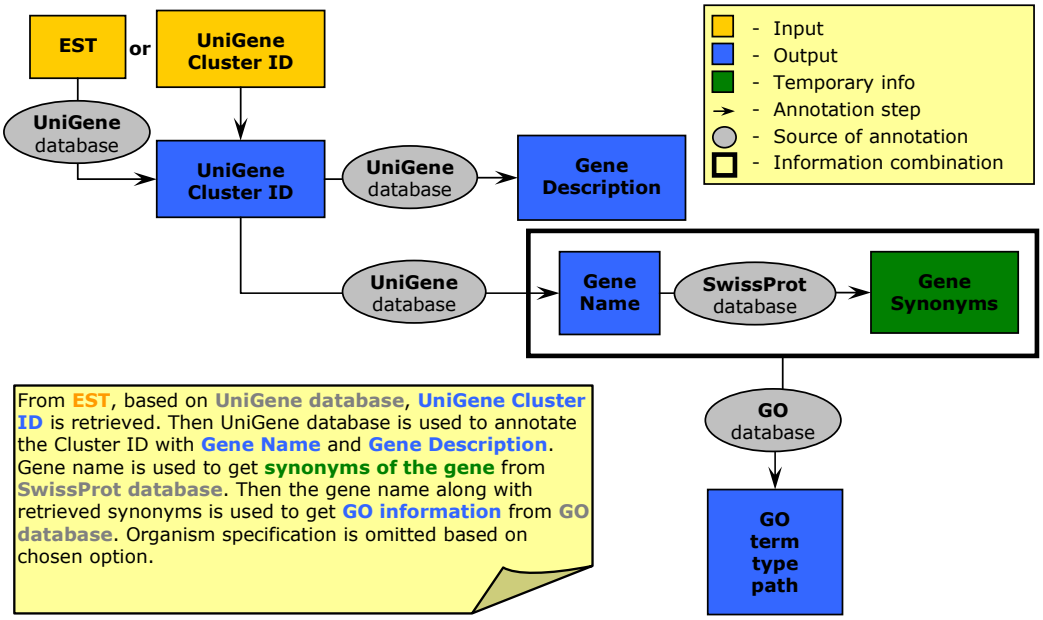

Figure 23. Schema of UniGene annotation plan.

Transfac annotation plan annotates each gene with transcription factors that regulate that gene. A parameter ‘evidence quality’ is available for specifying to use for instance when only experimentally confirmed transcription factor information is required. The plan is used to supply modified Bayesian Decomposition with prior coregulation information by grouping together genes that have the same transcription factor and therefore expected to be co-expressed under specific conditions.

Specific annotation plans called agents were implemented and used to download information from remote sources and put the data into local database in order to provide quick and reliable access to necessary data for many other annotation plans.

The list of agents is presented in Table 4. Those agents are currently installed to the ASAP system and are automatically run by ASAP in specific time intervals as indicated in the table to keep the data up to date. The complete list of annotation plans of ASAP that use information retrieved by agents are shown in Table 5.

Table 4. Agents: annotation plans that acquire information from remote sources and store it locally.

Agent name	Data description	Update time
UG	Unigene IDs, gene names and descriptions for Genbank accessions	2 weeks
SP	Gene symbols synonyms information from Swissprot	1 month
AF	Affymetrix microarray slides information	3 months
AG	Agilent microarray slides information	3 months
GO	Gene ontology information for genes	1 month
TF	Transcription factors for genes from TRANSFAC	Manual update

Table 5. Annotation plans of the ASAP system

Plan	Description	Agent used	Plans used
Unigene	Retreives cluster ID, gene name and description for genbank accession number	UG, SP	
Agilent	Retreives genebank accession number for Agilent slide probes	AG	Unigene
Affymetrix	Retreives genebank accession number for Affymetrix slide probes	AF	Unigene
Accessions	Retrieves cluster ID, gene name and description for different accession numbers		Agilent, Affymetrix, Unigene
Ontology	Retrieves gene ontology information for any accession number	GO	Accessions
Transfac	Retrieves transcription factors for any accession number	TF	Accessions

CHAPTER 5: ANALYSIS OF GASTROINTESTINAL STROMAL TUMORS DATA WITH MODIFIED BAYESIAN DECOMPOSITION

5.1 Data Analysis

5.1.1 Gastrointestinal Stromal Tumors

Until recently, there was no effective therapy for advanced, unresectable GISTs. Standard sarcoma therapies applied to the patients provided from little to no efficacy. A new agent, Gleevec (imatinib, mesylate, STI-571) has been shown to provide a significant classic response rate and the majority of patients that were treated with Gleevec demonstrated clinical improvement. However, there are cases when patients report no response to the treatment with little understanding of the mechanism.

An ongoing clinical trial at the Fox Chase Cancer Center focused on the effect of Gleevec on Gastrointestinal Stromal Tumors (GISTs) recently generated pre- and post- treatment samples (53 samples from 25 patients). The experiment is aimed at understanding of mechanisms of the disease, and in particular the reason that some patients are nonresponders to the treatment. Since imatinib mesylate is known to interrupt aberrant signaling in mutated c-KIT (the primary cause of most GIST), the lack of response is not well understood. The goal of this research is determination of the reason for treatment failure, whether rescue of the c-KIT pathway downstream of c-KIT or activation of other pathways. Identification of the key failure mode may identify new therapeutic targets or suggest additional therapeutic approaches.

5.1.2 Data from tumors and biopsies

RTOG S-0132 is a phase II NCI/CTEP approved clinical trial of neoadjuvant/adjuvant STI-571 (GLEEVEC NSC #716051) for primary and recurrent operable malignant GIST expressing the KIT receptor tyrosine kinase (CD117). Patients of 18 years and above, both genders, diagnosed with GIST (biopsy-proven) were recruited for the trial. Additional recruitment criteria included immunohistochemical documentation of KIT expression in the tumor and no history of chemotherapy, radiation therapy, biologic therapy, prior Gleevec. Or other investigational drug within 28 days of study entry.

Protocol of the study implied availability of a biopsy sample of sample before registering. The core specimens were obtained with the use of ultrasound, CT scan, or endoscopic guidance to assure adequate specimen retrieval in a nonnecrotic area of tumor. Patients started taking Gleevec within two weeks following registration. Patients stopped protocol treatment if their disease progressed at any time and were considered for surgery. The rest of patients underwent surgery after eight weeks of starting Gleevec treatment. Default Gleevec dose were set to 600mg per day and were modified based on toxicity grades for patients to 400mg (grade 2) or 200mg (grade 3 or 4) per day. Patients stopped Gleevec the night before surgery and underwent standard surgical resection with the objective of surgical debulking and attempt to remove all gross disease.

Experimental samples include biopsies of tumors taken from patients before treatment with Gleevec and a portion of the tumor received after dissection surgery

(after 3-8 weeks of the treatment started). A total of 50 patients were recruited for the study, however only for 25 generated both pre- and post- treatment microarray data due to failure to progress to surgery. There are several types of information available for each patient: site of tumor origin (stomach, small intestine, colon, or rectum); KIT immunohistochemical staining (positive for GISTs); tumor size; mitosis count (assessment of malignant potential); histology (spindle, epithelioid, or mixed); and risk (high or low, based on tumor size, mitosis count and histology).

Tumor sizes before (TSB) and after (TSA) treatment were taken to determine a relative tumor growth (TG), i.e

$$TG = \frac{TSA - TSB}{TSB} \tag{5.1.1}$$

Sorted by the tumor growth, patient response data is presented in Figure 24. Patients with less than 25% tumor reduction were assigned to non-responders group. Although general threshold for clinical response is 30%, this is based on measurement after 12 weeks on therapy. Since the time on therapy in this study was only 3-8 weeks, we believe that 25% is more appropriate for such a short time span. Another observation that favored our choice of threshold is that the first break in the graph of responses happens at 25%.

The protocol for sample preparation for the microarray experiment is shown in Figure 25. RNA from both pre- and post- treatment samples was isolated according to the method of Chomczynski and Sacchi [164] with modifications. The quality and quantity of the total RNA was checked using RNA Nano LabChip® (Agilent Technologies) according to the protocol provided. Some RNA samples were purified

and DNase treated with RNeasy Micro Kit (Qiagen). 50ng of RNA from the sample as well as Human Universal Reference RNA (HUR), Stratagene [165] were amplified with Ovation Aminoallyl RNA Amplification and Labeling System (NuGEN Technologies, INC). Amplification products were purified with NucleoSpin Extract 2 kit, then RNA was dried, dissolved in coupling buffer (0.1M NaHCO3, pH 9.0), and stained with Alexa Fluor 555 (HUR) or Alexa Fluor 647 (RNA from patient samples) 1 hr at room temperature in the dark. Then this coupling reaction was purified with QIAquick PCR purification Kit (Qiagen). 60pmol of patient samples versus 60pmal HUR were hybridized on Agilent Human Whole Genome Oligo 44K slides (GE2_44k_1005), washed using stabilization and drying solution, and scanned with G2565BA scanner (Agilent Technologies), according to manufacture specifications. In total, 53 microarray hybridizations have been done, covering 25 patients with available pre- and post- samples, with 2 patients with replicated post-treatment samples.

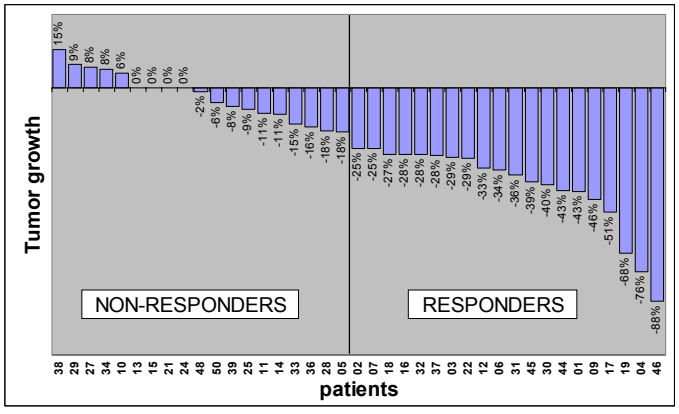

Figure 24. Relative tumor growth values. Percent of tumor growth is presented for patients with both pre- and post- treatment tumor sizes available. Patient was assigned to non-responders group if tumor reduced for less than 25%.

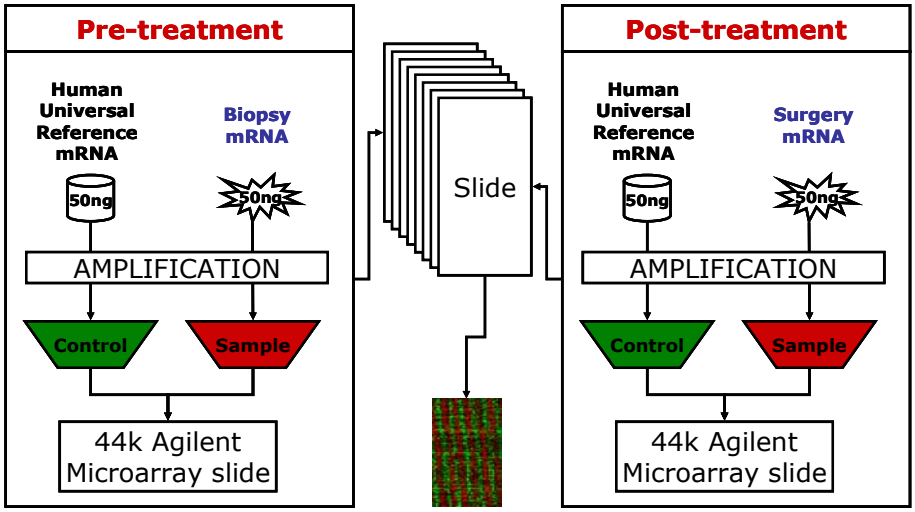

Figure 25. Preparation of samples for microarray experiment. 50ng of pre- and post- treatment samples were amplified and colored with Alexa Fluor 647 (red) and Human Universal Reference mRNA amplified and colored with Alexa Fluor 555 (green). Samples were hybridized on 44k human Agilent microarray slides and scanned with Agilent feature extraction software.

5.1.3 Preprocessing of GIST data

Microarray data provided after scanning of microarray slides was preprocessed in several steps. First, we reviewed scatter plots, where expression levels for each gene from two channels are plotted against each other. We found unexpected irregularities that are shown on plot as two distinct branches (Figure 26, center and right plots) in a cloud of expression points. Data from such microarrays were not considered for further analysis. After this filtration step, 16 patients had microarray data and only 8 of them had both pre- and post- treatment measurements available, resulting in total of 24 samples available for the analysis.

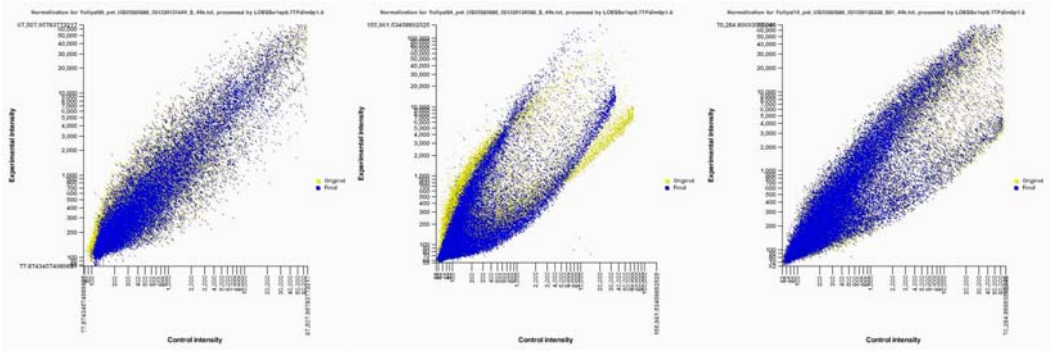

Figure 26. Example of scatter plots for microarray data from GIST patients. Leftmost figure represent expected distribution of expression values. Central and rightmost plots show bifurcation artifact due to errors during microarray experiment.

Data from microarray slides that showed no artifacts were normalized using LOWESS [60], with smoothing parameter of 0.7 and no correction for background. Probes with saturated signals in at least one channel were removed from slide and were not considered with the normalization method. Normalized data from different microarray slides was merged together and each probe was assigned with an expression value of ratio of experiment channel over control channel. Probes that were removed due to saturation in one or two channels were removed from the analysis.

Next preprocessing step included combining probes that are designed for the same sequence, i.e. combine replicates based on GenBank accession numbers associated with probes. The mean of expression ratios were taken as a signal and standard deviation of the mean as an uncertainty measurement associated with the signal. Probes without replicates received an uncertainty measurement of 20% of signal, a value based on average uncertainty measurements of replicated probes.

After removing probes that had saturated expression signals and combining probes linked to the same accession number together, we obtained a data matrix with measurements for genes corresponding to 33029 unique accession numbers.

5.1.4 Annotations for GIST data

We used the ASAP system to perform several annotations required for data filtering, composition and creation of prior coregulation groups for modified Bayesian Decomposition. Transcription factors and gene ontology information also was received with ASAP for interpretation of microarray data. GenBank accession numbers were used as primal identifiers for sequences represented on microarray slides to avoid poor reliability of gene symbols.

In order to compose data sets for Bayesian Decomposition analysis, we performed annotations for all 33029 accession numbers to receive corresponding gene information. The UniGene annotation plan (Figure 23) was used from the ASAP system to link each accession number to a UniGene cluster, gene name and description. RefSeq (EntrezGene) IDs would be more reliable, however the number of such genes is limited. For further analysis we retained 11733 sequences that were found in UniGene database to be linked to a known gene. By this we ensure that we analyze signals from sequences that truly represent genes which encode proteins and therefore that can be related to biological processes.

We applied the annotation plan that uses the professional TRANSFAC database to retrieve information about known transcription factors that regulate genes represented by accession numbers. A total of 988 genes from the 11733 UniGene

IDs had information (with the evidence of any quality) available in TRANSFAC at the time of the annotation. To generate prior coregulation groups we refined our search to pull out information with an evidence quality of 3 or higher, which includes transcription factors with functionally confirmed factor binding site (quality 1), binding of pure protein (quality 2), and immunologically characterized binding activity of a cellular extract (quality 3). Sequence based predictions therefore were not included in forming coregulation groups. Table 6 shows 69 groups of genes based on common transcription factor that have more than ten members in each. Among 631 unique genes that formed these groups, only 311 were represented in a single unique group, and others were regulated by more than one transcription factor. The amount of such genes reflects present knowledge about multiple regulations, and even those 311 single grouped genes are likely to have multiple transcription factors. We used 69 coregulation groups as prior information to include into analysis by Bayesian Decomposition.

5.1.5 Dataset composition

There are several important questions that can potentially be addressed by analysis of the described GIST data. First, genes that are biological markers of patient response to Gleevec treatment is of a special interest, since they can be used as a predictors of treatment efficacy. Second, analysis of the GIST microarray data can determine activity of transcription factors that differ between responders and non-responders. While biological markers can provide us with indicators of the disease, estimations of the activity of transcription factors can help in determining activity of

upstream signalling pathways and therefore in revealing of mechanisms of resistance to the treatment. To address these different questions we created two separate data sets that are designed to take advantage of the data in order to isolate the problems of interest.

Table 6. Coregulation groups based on transcription factors for GIST data set.

#	TF	Total genes	Shared genes	#	TF	Total genes	Shared genes	#	TF	Total genes	Shared genes
1	STAT5B	10	10	24	HMG I	13	13	47	NF-IL6-2	20	14
2	RAR-β	10	9	25	STAT5A	13	11	48	COUP-TF1	21	20
3	Nkx2-1	10	7	26	NF-AT1	13	12	49	HNF-1α-A	21	15
4	AhR:Arnt	10	2	27	IRF-1	14	6	50	COUP-TF2	21	20
5	GATA-6	10	9	28	p50	14	14	51	YY1	21	18
6	POU2F1	10	7	29	c-Myc	14	6	52	Gfi1	21	10
7	IPF1	10	7	30	c-Ets-1	14	8	53	RXR-α	22	19
8	NF-YB	10	9	31	ATF-2	15	12	54	GR	22	16
9	LXR-α:RXR-α	10	7	32	STAT1	15	14	55	HNF-3β	23	23
10	HSF2	11	11	33	GATA-1	16	3	56	HIF-1	24	12
11	NRSF	11	6	34	NF-YA	16	13	57	Egr-1	24	20
12	MyoD	11	6	35	GATA-4	16	14	58	ER-α	26	18
13	GLI1	11	4	36	Pax-6	16	2	59	HNF-4α1	28	26
14	HSF1 (long)	12	12	37	C/EBPδ	17	17	60	USF1	29	26
15	HNF-6α	12	12	38	STAT3	17	12	61	NF-κB	29	26
16	HNF-3α	12	12	39	LEF-1	17	7	62	HNF-4	29	21
17	E2F-1	12	7	40	c-Fos	18	18	63	SRF	31	24
18	HNF-6β	12	12	41	C/EBPβ	18	18	64	c-Jun	38	36
19	T3R-α	12	8	42	NF-Y	19	13	65	Sp3	45	37
20	T3R-β1	12	10	43	AP-2αA	19	17	66	CREB	45	32
21	JunD	12	12	44	HNF-4α	19	13	67	AP-1	48	41
22	E12	12	9	45	RelA	19	19	68	C/EBPα	52	47
23	JunB	12	12	46	USF2	20	19	69	p53	71	24

The first data set (named the biomarkers data set) comprised genes that were significantly differentially expressed between responders and non-responders, where threshold for non-responders group was defined as shown in Figure 24. We applied significance analysis of microarrays (SAM) [80] to the expression data of 11733 annotated genes and received a list of 581 genes that were significantly different between two patient groups with a false discovery rate of 13%. Out of pre- and post-

treatment samples of 16 patients, samples of patients 10, 13, 28, 29, 34, 41 and 50 were in the group of non-responders, samples of patients 01, 03, 09, 12, 19, 22 and 45 were in the group of responders, and samples of patients 08 and 23 were unclassified, since no data about tumour size after treatment were available for these patients at the time of the analysis. The final data matrix consisted of measurements of 581 genes across 24 samples.

The second data set (named the transfac data set) comprised 988 genes that had annotations from the professional TRANSFAC database with any evidence quality. Designed to determine activity of transcription factors and possibly of an upstream signalling pathway, the final data matrix comprised expression levels of 988 genes across 24 samples.

5.1.6 BD analysis

We analyzed the biomarkers dataset using Bayesian Decomposition without any prior coregulation information included in the analysis, since it was not enough information of verified transcription factors for 581 genes that were included in the data set. 9 separate Bayesian Decomposition instances were run positing from 2 to 10 patterns to cover the possible dimensionality of the data. The transfac data set was analyzed with modified Bayesian Decomposition using coregulation groups from Table 6 as prior information for the analysis. BD was run positing from 5 to 15 patterns, covering more solutions than for the biomarkers dataset to account for patterns that do not relate to patient response, since the data were not filtered based on expression profiles of the included genes and therefore contained more diversity.

5.2 Results

5.2.1 Methods of result interpretation

There are several steps in interpretation of the results received with Bayesian Decomposition. First, an issue of determining a correct number of patterns in the analyzed data requires a solution. Once the number of patterns is determined, amplitude and pattern matrices should be analyzed to reveal patterns that represent meaningful expression changes and groups of genes that belong to such patterns. Then, groups of genes are investigated to see if they possess any distinctive features, for example, by comparing portion of genes from the group involved in a certain biological process to the portion of such genes from the whole dataset (refer to equation 5.2.1).

The problem of defining a correct number of expression patterns that exist in the data can be addressed by looking at the persistence of patterns across results with different numbers of patterns posited in the analysis. The persistence of a pattern can be defined as the number of consecutive times that a pattern is found by BD as we increase the number of posited patterns. Therefore, one would expect average persistence to decline monotonously with increased number of patterns. A point, after which monotony is broken by sudden drop in average persistence of found patterns, would indicate an optimal number of expression patterns in the data. That is, giving unneeded degree of freedom to the model results in a set of patterns that lack consistency.

The goal of the GIST data analysis focuses on patient response to Gleevec treatment. Therefore, the main criterion of interest for patterns is a correlation of gene expression with observed response. Other possible biologically meaningful patterns include correlation with patient's gender, age, origin of a tumor, etc.

The amplitude matrix can be used to see what genes contain patterns of interest and group those genes together to perform enhancement analysis of the group. Enhancement of a term in a group is a measure of presence of the members annotated with the term in the group compared to presence of the term in the whole data, i.e.:

$$E_{term} = \frac{N_{term}^{group} / N_{total}^{group}}{N_{term}^{data} / N_{total}^{data}} \tag{5.2.1}$$

where N_{term} is a number of genes annotated with the term in the group or in the whole data and N_{total} is a number of all genes in the group or in the whole data. Thus, enhancement can provide interpretation for a group of found co-expressed genes and assign biological meaning to the expression pattern by monitoring properties of genes that contain the pattern.

5.2.2 Biomarkers data set

We analyzed the biomarkers data set using Bayesian Decomposition without prior coregulation information included into analysis. First, we calculated average persistences for different number of solutions (patterns) posited into analysis. Based on the graph shown in Figure 27, we picked 7 patterns as the optimal number of expression patterns for analysis of the data.

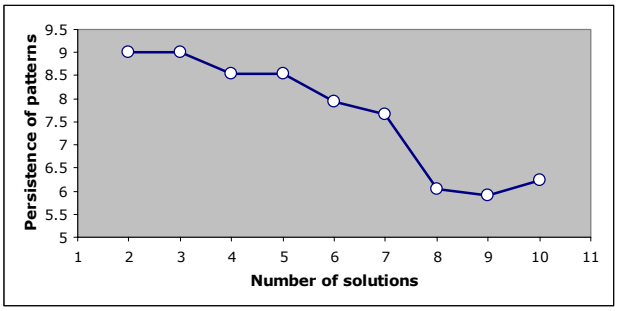

Figure 27. Average persistences of patterns for different number of solutions posited into analysis in biomarkers data set. When increasing number of patterns from 7 to 8, the curve drops faster than expected indicating that increasing number of solution after 7 results in relatively non-stable patterns.

All 7 expression patterns from the pattern matrix recovered by Bayesian Decomposition were used to calculate correlation coefficients with patient response. Two patterns with absolute correlation coefficients >0.7 were found, pattern 4 being positively correlated with response with value of correlation R=0.702 and pattern 7 being negatively correlated with response with value of correlation R=-0.786. Figure 28 shows those patterns. Based on the expression values related to unclassified patients 08 and 23 in recovered patterns linked to patients' response, we concluded that both patients 08 and 23 should belong group of responders. We will use these predictions as a validation point of our analysis, when true information about the response is received.

The Amplitude matrix from the Bayesian Decomposition analysis was used to assign genes to patterns 4 and 7. Each gene in the amplitude matrix had a mean value of its strength within a pattern and an uncertainty on that assignment based on the MCMC sampling performed by BD. We used a threshold of 3 standard deviations away from zero as a requirement to assign a gene to a pattern, which resulted in a group of 194 genes that contained pattern 4 and 46 genes that contained pattern 7.

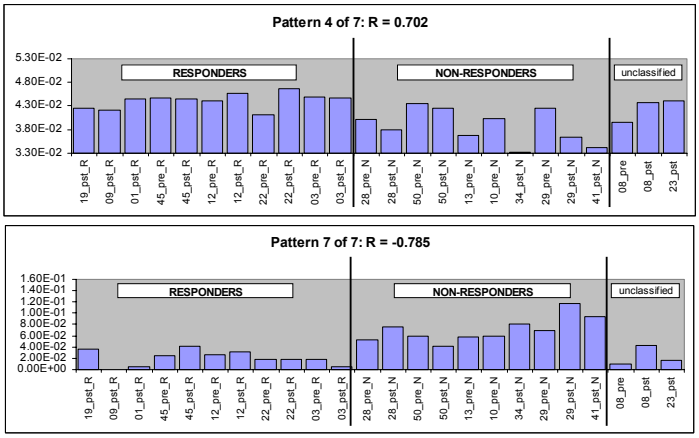

Figure 28. Expression patterns found by Bayesian Decomposition in biomarkers data set. Figure on the top shows pattern positively correlated (R=0.702) with response (genes are upregulated in responders compared to non-responders). Figure on the bottom shows pattern negatively correlated (R=-0.785) with response (genes with such expression pattern are downregulated in responders compared to non-responders).

We performed enhancement analysis of gene ontology terms for these groups of genes based on annotations from the ASAP system. Table 7 and Table 8 represent gene ontology enhancements of 1.5 and above for the groups of genes that contain expression pattern 4 and pattern 7 respectively.

Table 7. Table of gene ontology term enhancements for the pattern correlated with response.

Gene Ontology Biological Process	**Enhancement**	N_{term}^{group}	N_{term}^{total}
GO:0006935 chemotaxis	*2.62*	*7*	*8*
GO:0042330 taxis	2.62	7	8
GO:0042221 response to chemical substance	2.00	8	12
GO:0016337 cell-cell adhesion	*1.91*	*7*	*11*
GO:0001501 skeletal development	1.87	10	16
GO:0019725 cell homeostasis	1.87	5	8
GO:0006118 electron transport	1.83	11	18
GO:0042592 homeostasis	1.66	5	9
GO:0006643 membrane lipid metabolism	1.66	5	9
GO:0009628 response to abiotic stimulus	1.65	11	20
GO:0030029 actin filament-based process	1.63	6	11
GO:0030036 actin cytoskeleton organization and biogenesis	1.63	6	11
GO:0043067 regulation of programmed cell death	*1.61*	*7*	*13*
GO:0042981 regulation of apoptosis	*1.61*	*7*	*13*
GO:0000004 biological process unknown	1.59	9	17
GO:0006091 generation of precursor metabolites and energy	1.56	13	25
GO:0007155 cell adhesion	*1.54*	*20*	*39*

Table 8. Table of gene ontology term enhancements for the pattern correlated with nonresponse.

Gene Ontology Biological Process	**Enhancement**	N_{term}^{group}	N_{term}^{total}
GO:0042127 regulation of cell proliferation	*3.16*	*5*	*20*
GO:0048522 positive regulation of cellular process	2.44	6	31
GO:0051242 positive regulation of cellular physiological process	2.43	5	26
GO:0008283 cell proliferation	*2.39*	*7*	*37*
GO:0048518 positive regulation of biological process	2.37	6	32
GO:0043119 positive regulation of physiological process	2.34	5	27
GO:0043118 negative regulation of physiological process	1.97	5	32
GO:0051243 negative regulation of cellular physiological process	1.97	5	32
GO:0007165 signal transduction	1.90	19	126
GO:0007242 intracellular signaling cascade	1.80	6	42
GO:0048523 negative regulation of cellular process	1.80	5	35
GO:0048519 negative regulation of biological process	1.75	5	36
GO:0030154 cell differentiation	*1.75*	*5*	*36*
GO:0007166 cell surface receptor linked signal transduction	1.64	7	54
GO:0045184 establishment of protein localization	1.62	5	39
GO:0008104 protein localization	1.62	5	39

5.2.3 Transfac data set

The transfac data set was analyzed using modified Bayesian Decomposition with prior coregulation information from Table 6 included into analysis. Similar to biomarkers data set, we calculated average persistences for different number of solutions posited into analysis and average persistence plot dropped after 9 patterns as demonstrated in Figure 29. After this step we analyzed pattern and amplitude matrices recovered by modified BD with 9 patterns posited into decomposition process.

Two basic steps were performed for each pattern and corresponding group of genes that contained the pattern. First, 3 correlation values of a pattern with response of only pre- treatment samples, only post-treatment samples and all samples were calculated. Second, we performed transcription factors and gene ontology term enhancement analysis for all groups, using ASAP annotations from the professional TRANSFAC database and the Gene Ontology Consortium database. Only

information that had qualities of evidence of 1, 2 and 3 were used for transcription factor enhancement analysis. A gene was assigned to a pattern if corresponding mean value (sampled during MCMC process) from amplitude matrix satisfied the condition to be of 3 standard deviations above zero.

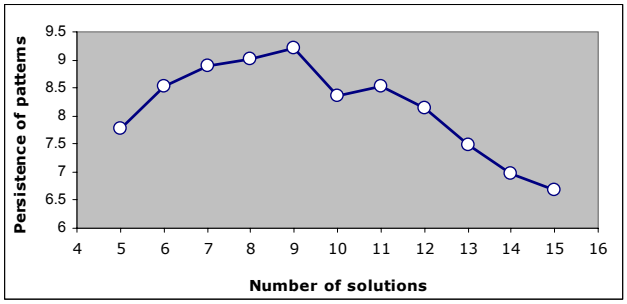

Figure 29. Average persistences of patterns for different number of solutions posited into analysis in Transfac data set.

Although correlations calculated for all samples of nine patterns received from analysis of the Transfac data set didn't indicate strong link with the response as can be seen in column 4 of Table 9, we decided to focus on specific patterns that indicate connection to response for specific pre- or post- treatment samples (columns 2 and 3 of Table 9). Thus, patterns with positive and negative correlation values of magnitude 0.4 and above were chosen for further analysis. Correlation of a pattern's pre-treatment samples with response can indicate differences in activity of transcription factors and upstream signalling pathways between two groups of patients before starting therapy. Markers of response can help in identifying of underlying mechanisms lying behind the difference. Similarly, patterns with post-treatment samples linked to outcome can be treated as biological processes

differentially expressed between responders and non-responders due to introduction of Gleevec into the system. Potentially these can help in identifying signalling pathways that lead to rescue of cancer cells in non-responsive patients, providing additional therapeutic targets.

Table 9. Correlation with response for patterns found in Transfac dataset. Different columns show correlations for specific expression values from patterns that correspond to different sample types. Correlations of absolute value >0.4 are marked in bold italic.

Pattern	Only pre-treatment samples	Only post-treatment samples	Both, pre- and post-treatment samples
1 of 9	-0.01	-0.35	-0.19
2 of 9	-0.26	***0.54***	0.38
3 of 9	0.01	0.16	0.14
4 of 9	***0.47***	0.07	0.21
5 of 9	-0.05	***-0.41***	-0.23
6 of 9	0.12	0.20	0.09
7 of 9	-0.27	***0.41***	-0.14
8 of 9	-0.12	-0.29	-0.21
9 of 9	***-0.58***	-0.04	-0.15

Table 10 demonstrates enhancement analysis results for the patterns 2,4,5,7 and 9 that showed correlations of pre- or post- samples with response. In order to avoid false positive errors, results were filtered to include only such terms that are represented within 4 or more members of a pattern and showed two-fold enhancement. Results of enhancement analysis of associated with genes transcription factors are reported in Table 11 (pattern 7 did not show significant enhancements). Even more stringent filtering criteria were applied to the results. Only greater than two-fold enhanced transcription factors are reported with 4 or more members presented in the group. Plus, threshold of 5% was chosen as a maximum probability for reported transcription factor to be found randomly. Probabilities of error were

calculated based on hypergeometric distribution, i.e. probability P of grouping together exactly x of a possible K genes annotated with a given transcription factors in a group of M genes from total number of N genes (988 for the Transfac data set), i.e.

$$P(x) = \frac{\binom{K}{x}\binom{N-K}{M-x}}{\binom{N}{M}} \tag{5.2.2}$$

Table 10. Enhancements of biological process GO terms for groups of genes with selected from Transfac data set patterns.

Gene Ontology term	**Enhancement**	N_{term}^{group}	N_{term}^{total}
Pattern 2 (for post-treatment samples correlation with response R=0.54), 129 genes			
GO:0044257 - cellular protein catabolism	2.55	4	12
Pattern 4 (for pre-treatment samples correlation with response R=0.55), 78 genes			
GO:0006366 - transcription from RNA polymerase II promoter	5.43	6	14
GO:0008285 - negative regulation of cell proliferation	2.53	5	25
Pattern 5 (for post-treatment samples correlation with response R=-0.41), 97 genes			
GO:0008610 - lipid biosynthesis	2.72	4	15
GO:0006811 - ion transport	2.04	5	25
GO:0006066 - alcohol metabolism	2.04	4	20
Pattern 7 (for post-treatment samples correlation with response R=0.41), 63 genes			
GO:0009628 - response to abiotic stimulus	2.61	4	24
Pattern 9 (for pre-treatment samples correlation with response R=-0.58), 95 genes			
GO:0050790 - regulation of enzyme activity	4.16	4	10
GO:0006520 - amino acid metabolism	2.77	4	15
GO:0044249 - cellular biosynthesis	2.21	7	33
GO:0006519 - amino acid and derivative metabolism	2.17	5	24

Table 11. Enhancements of transcription factors for groups of genes with selected from Transfac data set patterns.

Transcription Factor term	Enhancement	N_{term}^{group}	N_{term}^{total}	Random?
Pattern 2 (for post-treatment samples correlation with response R=0.54), 129 genes				
KCHIP2.6	6.13	4	5	0.122%
CSEN	6.13	4	5	0.122%
c-Jun:c-Fos	5.96	7	9	0.002%
STAT3	4.25	5	9	0.260%
HMG I	3.48	5	11	0.727%
AP-1	3.48	5	11	0.727%
LXR-alpha:RXR-alpha	3.4	4	9	1.780%
NF-Y	3.4	4	9	1.780%
STAT1	3.06	6	15	0.667%
p50	2.55	4	12	4.638%
c-Jun	2.14	7	25	2.411%
SRF	2.09	6	22	3.851%
Pattern 4 (for pre-treatment samples correlation with response R=0.55), 78 genes				
VDR	7.24	4	7	0.100%
STAT3	5.63	4	9	0.308%
STAT1	4.22	5	15	0.376%
RXR-alpha	4	6	19	0.202%
c-Myc	3.9	4	13	1.275%
SRF	3.38	4	15	2.078%
p53	2.49	12	61	0.132%
Pattern 5 (for post-treatment samples correlation with response R=-0.41), 97 genes				
Max1	10.19	5	5	0.001%
STAT3	4.53	4	9	0.673%
c-Myc	3.92	5	13	0.485%
NF-Y	3.4	4	12	1.958%
p53	2.91	4	14	3.241%
RXR-alpha	2.68	5	19	2.415%
Gfi1	2.55	5	20	2.914%
Pattern 9 (for pre-treatment samples correlation with response R=-0.58), 95 genes				
AFP1	10.4	4	4	0.008%
HNF-1alpha-B	10.4	6	6	0.000%
HNF-1alpha-C	10.4	6	6	0.000%
c-Jun:c-Fos	6.93	6	9	0.004%
HMG I	4.73	5	11	0.193%
HNF-4alpha	4.33	5	12	0.301%
AP-1	3.78	4	11	1.347%
HNF-1alpha-A	3.47	7	21	0.191%
HNF-3beta	3.31	7	22	0.255%
HNF-4	3.25	5	16	1.127%

5.2.4 Comparison of results recovered from Transfac data set with modified and original Bayesian Decomposition

Comparison of results of analysis for two algorithms without knowing correct answers is a nontrivial task. We ran original BD positing 9 patterns into analysis

without including any additional knowledge. Similarly to results of modified BD, correlation coefficients of recovered patterns with response were calculated and enhancement analysis of transcription factors performed according to protocol described in Section 5.2.3.

All correlation coefficients with response for patterns received by original and modified BD were compared. Average absolute difference for all correlation coefficients as well as only for above threshold values (|correlation|>0.4) was less than 5% of average correlation magnitude for each of the methods, indicating insignificant difference between recovered expression patterns. However, amount of genes assigned to these patterns varied greatly, with modified Bayesian Decomposition assigned 462 genes to significant patterns 2, 4, 5, 7 and 9 (129, 78, 97, 63 and 95 genes respectively), while original Bayesian Decomposition had a statistical power to assign only 385 genes (90, 72, 94, 65 and 64 genes) to these patterns, a drop of 17%. It appears therefore that the modified version of BD is using prior information to increase the linking of genes to patterns, as desired. On contrary, insignificant patterns 1, 3, 6 and 8 were assigned with 225 by modified BD and 331 genes by the original BD.

Enhancement analysis performed for gene groups revealed consequences of the lack of statistical power to assign genes to a pattern for original Bayesian Decomposition. Pattern 4 did not indicate activity of VDF, and significance of enhancement of activity of the transcription factor Max1 was lost for Pattern 5. Additionally, we compared amount of genes related to significantly enhanced transcription factors (enhancement greater than 2.0, with 4 or more members

presented in the group) recovered from results received with and without using prior coregulation information. Number of enhanced terms, grouped and total number of genes with these terms for significant patterns 2, 4, 5, 7 and 9 for both methods are shown in Table 12. The table accounts for more terms than Table 11, since it also contains terms that do not satisfy 5% error rate threshold. Those results indicate that using coregulation information to overcome the low signal-to-noise ratio of microarray data on average helped in recovering of ~34% of group members, while algorithm without using such information produced coverage of only ~24%.

Table 12. Comparison of number of genes with enhanced terms from results received with and without using coregulation information.

	Using coregulation			**Without coregulation**		
	Number of enhanced terms	Genes with the terms in the group	Genes with the terms in the data	Number of enhanced terms	Genes with the terms in the group	Genes with the terms in the data
Pattern 2	14	69	171	11	52	206
Pattern 4	9	47	183	2	13	74
Pattern 5	9	40	130	1	4	20
Pattern 7	0	0	0	0	0	0
Pattern 9	12	68	177	4	18	56
Average recovery of genes		34%			24%	

5.2.5 Discussion

Two expression patterns found in the biomarkers data set were linked to patient outcome. Pattern 4, which showed positive correlation with response, included 194 genes. The list of 42 genes with >90% of their behavior explained solely by pattern 4 was investigated further to look for connections with previous studies. Figure 30 shows an amplification of chromosomal region 6p21 found in an independent study

of DNA mutations in GIST patients by Dr. Godwin's lab at the Fox Chase Cancer Center. Chromosome locations of 3 genes, SRF, MAD2L1 and VEGF, with respectively 97%, 91% and 95% expression profile explained by pattern 4, lie in the region. Further inspection of the list of genes with expression pattern 4 yielded one more gene, BYSL, which resides near the region. Unfortunately, because no outcome data was available for the patients that show such DNA mutation, we could not validate importance of these four genes in prediction of the response.

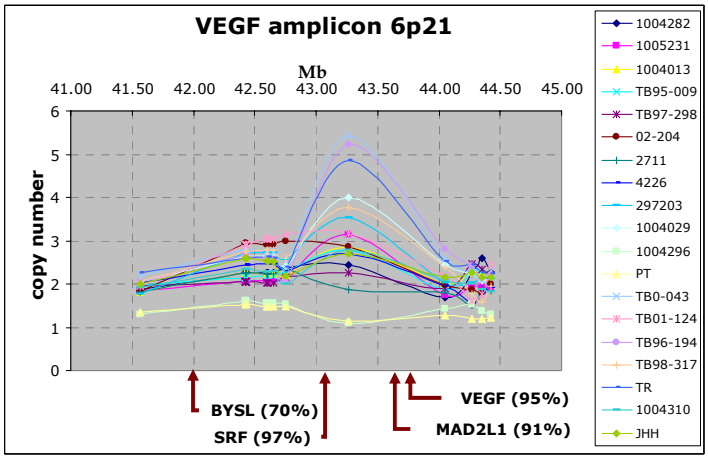

Figure 30. Chromosome copy analysis of cytoband 6p21 for GIST patients from different study. Common amplification region in shown on the figure. BYSL, SRF, MAD2L1, VEGF – genes than have expression profiles explained by pattern 4 for 70%, 97%, 91% and 95% respectively.

Gene ontology enhancement analysis of the genes that contain pattern 4 and pattern 7 was performed to search for biological significance of the patterns. One of several significant enhancements for pattern 4, presented in Table 7, included cell adhesion related biological processes that are known to be mutated in some cancers, resulting in abnormal cell-to-cell interactions and tumor growth. Another difference

between responders and non-responders was in regulation of apoptosis, most probably related to higher apoptosis after Gleevec treatment, responsible for the observed difference in patient outcomes. Gene ontology enhancements for the group of genes with expression pattern 7, which is negatively correlated with response, also followed the prediction model. Marked with italic font in Table 8, cell proliferation and differentiation processes are overrepresented in non-responders compared to responders, indicating tumor growth or smaller size reduction of tumors in patients that do not respond to the imatnib.

After analysis of the biomarkers data set we received missing data for post-treatment tumor sizes of patients 08 and 23. The data for patients 08 and 23 showed response of 39% and 27% tumor reduction size respectively. Thus, the prediction we made based on the relative expression levels for samples 08_pre, 08_pst and 23_pst in both pattern 4 and pattern 7, validated successfully, providing confirmation of prediction of response made by these patterns.

Analysis of the genes with known transcription factors revealed expression patterns that were linked to the response based on correlations of only pre- or post-treatment samples with outcome. From nine patterns, with the number of patterns determined by analysis of average pattern persistences, five showed correlations with tumor reduction. 4 of these patterns, namely pattern 2, pattern 4, pattern 5 and pattern 9 showed significant (hypergeometric based error rate <5%) enhancements of several transcription factors as demonstrated in Table 11.

Pattern 2 showed positive correlation with response for post-treatment samples (R=0.54). Significant enhancements of gene ontology terms included only non-

informative biological processes of a metabolic nature. Activity of such transcription factors as KCHIP2.6 (also called FOSL2), which had an enhancement of 6.13, belongs to family of FOS genes and encode leucine zipper proteins that can dimerize with proteins of the JUN family, forming the transcription factor complex AP-1. Activity of AP-1 was also determined for the pattern2, along with the c-Jun:c-Fos complex, which is another instance of AP-1 transcription factor. The activity of these transcription factors demonstrate cancer-related biological processes as FOS proteins have been implicated as regulators of cell proliferation, differentiation, and transformation.

Another group of transcription factors activated according to expression pattern 2 that reflects response in post-treatment samples consist of serum response factor (SRF) and STAT1/STAT3 genes, also represented in pattern 4 with pre-treatment samples correlated to response (R=0.55). It is known that both STAT1/STAT3 and SRF contribute to c-fos promoter activation [166] and thereby participates in cell cycle regulation, apoptosis, cell growth, and cell differentiation processes.

Some of the genes that contain pattern 4 are also regulated by vitamin D receptor (VDR) that showed enhancement of 7.24 and is known to inhibit growth of breast cancer cells [167, 168]. Since pre-treatment samples of pattern 4 positively correlated with response, we can predict that the activity of signalling pathway that regulates VDR potentially aids in inhibition of GIST cell growth after therapy with Gleevec. Another conclusion can be drawn for the pattern 4 from enhancement of negative regulation of cell proliferations gene ontology term (Table 10). It indicates presence

of genes that contribute to inhibition of cell proliferation in most responders, but not in non-responders.

Pattern 5 demonstrated negative correlation of post-treatment samples with response (R=-0.41), meaning that genes included in the group are more expressed in post-treatment samples of responders compared to non-responders. More than ten-fold enhancement of genes regulated by Max1 is given in Table 11 with 5 of total 5 genes from the data set grouped by Bayesian Decomposition. Max1 is a heterodimeric partner of c-Myc that allows self dimerization. While Myc-Max compound activates transcription to promote apoptosis, Max-Max may repress it due to lack of a transcriptional activation domain [169, 170]. Such behavior of Max1 is perfectly explained within the model – expression levels of Max1 genes that contain pattern 5 are higher in non-responders (post-treatment samples) indicating activity of Max1 that rescues GIST sells from apoptosis.

Pattern 9 with pre-treatment samples correlated negatively with response (R=-0.58) showed greater than ten-fold enhancements for two transcription factors, AFP1 and HNF-1alpha (target genes for HNF-1alpha-B and HNF-1alpha-C were the same, so do not allow separation of these three transcription factors). Both AFP1 and HNF1 are known to have important biological role in carcinomas, with AFP1 activity linked to hepatocellular carcinoma in liver [171, 172] and HNF-1alpha loss of function in renal cell carcinoma [173].

One of the results that bind biomarkers and transfac data together is a recovery of a transcription factor SRF in both analyses. Shown to be amplified in some GIST patients (Figure 30), SRF was assigned to the pattern 4 of 7 for the biomarkers data

set, which was positively correlated with response to Gleevec treatment. As expected, activity of the SRF transcription factor was confirmed during analysis of the transfac data set. Groups assigned to patterns 2 and 4, which were also positively correlated with response in pre- and post- treatment samples, demonstrated enhancement of the SRF transcription factor regulated genes in these groups. While the probability of the results to be a false positive error cannot be ignored, prediction validation and various confirmations of biological meaning of recovered patterns related to SRF strongly suggest importance of the gene in GISTs and provide an important validation checkpoint to our designed data mining approach.

CHAPTER 6: CONCLUSIONS AND REMARKS

6.1 Conclusions

This study introduced a data mining process for analysis of high-throughput biological data, initially microarray data, that allows the inclusion of prior biological information to allow data integration and reduce the impact of noisy data. The core technique is the Bayesian Decomposition algorithm. Bayesian Decomposition uses a Markov Chain Monte Carlo sampler with Bayesian statistics that provides a mechanism for encoding prior knowledge into the process. A modification for the algorithm has been implemented that enables using known information about gene coregulation in the model. Such coregulation information along with annotations for all genes that are also necessary in the data mining process can be found in various databases with web-based access, but collecting this information can be time consuming and involve multiple steps before the data is presented in a specific usable form. To automate the process of data collection we created the Automated Sequence Annotation Pipeline system that allows customized annotation of data.

The data mining process was applied to analysis of Gastrointestinal Stromal Tumor data as shown on Figure 31. The data set analysis focused on the effects of Gleevec treatment on patients and mechanisms of difference in response to the drug. ASAP system and modified Bayesian Decomposition were used as a part of the microarray data analysis process and results were interpreted with received annotation information about transcription factors and gene ontology of genes under study.

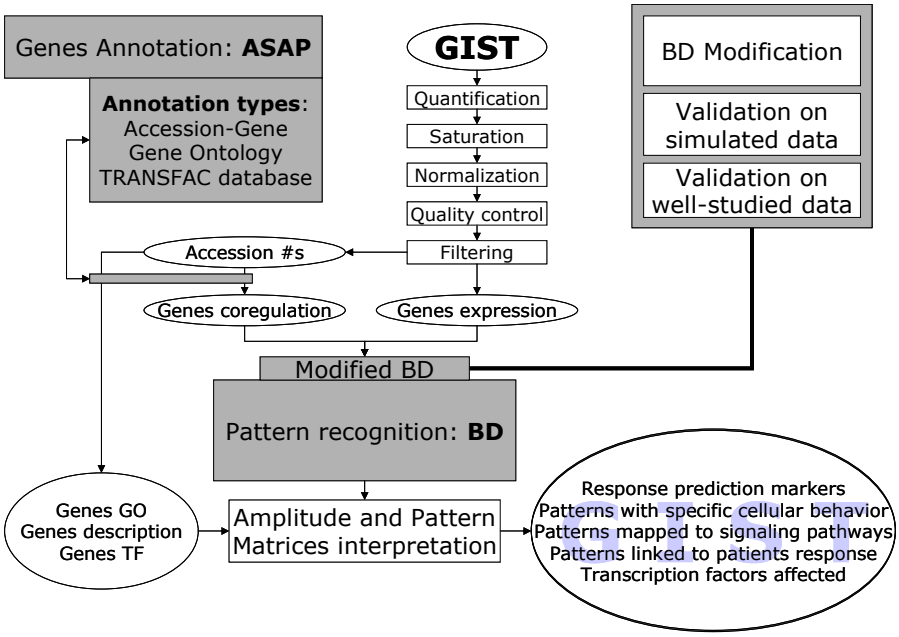

Figure 31. Summary of thesis contributions. GIST microarray data was pre-processed, annotated with ASAP system, analyzed by modified Bayesian Decomposition, which was validated on simulated and well-studied yeast data sets, and the results were interpreted using annotation information.

6.1.1 Modified Bayesian Decomposition

This thesis focused on creating a modified Bayesian Decomposition approach that allows encoding of prior coregulation information into analysis. We created convolution function that spreads atoms from one part of atomic domain into multiple amplitude matrix elements that correspond to a group of coregulated genes as described in Section 3.1. A normalization issue of genes having different expression levels was solved by applying additional step of Bayesian Decomposition analysis to determine weights for each gene in each group of co-expression (Figure 12).

Validation of modified Bayesian Decomposition was first performed on simulated data modeling yeast cell-cycle expression data. Multiple data sets were created with 154 different combinations of additive and multiplicative noise levels

that reflect low signal-to-noise ratio of real microarray measurements. In Section 3.2 we showed that including of prior knowledge into analysis improves quality of the received results. A significance of amount of such prior knowledge was also demonstrated by comparing results between different levels of prior included into analysis as shown in Figure 15.

The second step of validation of enhancements done to Bayesian Decomposition was done by analyzing real biological microarray data – yeast cell-cycle data set [90] and yeast mutant Rosetta Compendium data set [153]. ROC analysis (Section 3.2.1) was used for the purpose of comparing original and modified Bayesian Decomposition. Both data sets confirmed advantages of using prior coregulation information when compared by ROC analysis with a golden standard that comprises the same information (left Figure 17, Figure 18). A different golden standard based on groups of Cherepinky *et al*. [94], was used for ROC analysis of yeast cell-cycle results where there was no difference in recovering of the golden standard information between original and modified Bayesian Decomposition. This result demonstrates present limits on coregulation knowledge, since only 67 from 788 genes had prior information associated with them and that wasn't enough to provide a significant difference in the results. Nevertheless, we also compared Bayesian Decomposition algorithm to hierarchical and k-means clustering. The results of the comparison shown in Figure 17 (hierarchical clustering) and Figure 18 (k-means clustering) suggest that important ability of Bayesian Decomposition to allow assignment of one gene to multiple patterns of expression is superior to clustering approach of one gene to one cluster membership.

6.1.2 Automated Sequence Annotation Pipeline

The Automated Sequence Annotation Pipeline was created to allow customized annotation that can query various sources including remote and local databases, and take output from one query as the input for a new query (Chapter 4). User and administrator oriented web interface of the system was designed to simplify access of users to annotation plans of the system and received results and provide administrators with tools to manage users, annotation plans and messaging system that alarms about errors due to changes of format of data provided by independent sources of annotations. ASAP was implemented and installed in Fox Chase Cancer Center allowing access of researchers of the institution to aid in studies by providing various annotations.

Several annotation plans were designed to support the data mining process described in this thesis support at all steps of the analysis as depicted in Figure 4. Probe annotation plans are available for Agilent and Affymerix platforms allowing access to constantly updated microarray slides information for these systems. UniGene annotation plan was created for receiving cluster ID, gene name and description for GenBank accession numbers and used for analysis of GIST data set for filtering and data composition steps (Section 5.1.3). Gene ontology and transcription factors information for genes can be also received by annotation plans implemented within ASAP system. We successfully used these annotations to create prior information, presented in Table 6, for analysis of GIST data set with modified Bayesian Decomposition. Finally, enhancement analysis of the groups of genes

received from GIST data set was performed using gene ontology and transcription factors annotation information.

6.1.3 Analysis of Gastrointestinal Stromal Tumors

The central study of Gastrointestinal Stromal Tumors data was preformed according to the developed data mining process. Annotations received with ASAP system were used at different steps of the analysis as described in Section 5.1.3, including gene filtering, data composition, prior information generation and gene enhancement results analysis. Data composition step resulted in generating two separate data sets that were designed to carry out two different tasks: to find genes that can serve as biomarkers to predict patients response to Gleevec and to determine activity of transcription factors that can explain difference in such a response.

The biomarkers data set revealed genes that contain expression patterns correlated with patients' response. Two groups of 194 and 47 genes were found that had a correlation with response of 0.702 and -0.785 respectively. Gene ontology enhancements analysis was performed to see the biological processes that these genes were involved in. Positively correlated with response patterns showed biologically meaningful enhancements of cell apoptosis processes, indicating higher rate of cell deaths in responders to Gleevec treatment. On contrary, pattern that showed more expression in non-responders, revealed enhanced biological processes of cell proliferation and differentiation, which can be related to slower size reduction of tumor or even tumor growth in non-responsive patients. Finally, predictions made for

patients 08 and 23 to be responders were confirmed as correct with the later received missing outcome data.

The second data set revealed activity of various transcription factors that can be linked to patients response based on sample before and after Gleevec treatment. SRF transcription factor determined by the analysis of biomarkers dataset to be upregulated in responders compared to non-responders, also showed increased activity in responders in both pre- and post- treatment samples. Another transcription factor, Max1, showed activity enhancement of more than 10-fold with negative correlation to response in post-treated samples. Known to have and ability to rescue endothelial cells from apoptosis, Max1 can be involved in similar mechanisms in GISTs, allowing tumor grow after being introduced with Gleevec. Also, activity of vitamin D receptor, VDR, transcription factor was demonstrated on its target genes to be more active in responders than in non-responders in pre-treatment samples. Combined with information from other studies, where activation of VDR was shown to inhibit growth of breast cancer cells, we can hypothesize about positive role of VDR during treatment of GISTs with imatinib.

Transfac data set was also analyzed with Bayesian Decomposition without any prior information included into the model. Comparisons of results were performed and revealed that original Bayesian Decomposition recovered 17% less genes for linked to response patterns 2, 4, 5, 7 and 9 from the data and failed to identify activation of transcription factors Max1 and VDR, while additional prior knowledge helped modified Bayesian Decomposition to assign expression patterns to genes that were confirmed to contribute to the statistical power of found results. Thus,

enhancements done to Bayesian Decomposition were validated on human microarray data and demonstrated to improve significance of received results.

6.2 Future Studies and Prospects

Contaminations of sample RNA occurred during microarray experiment resulted in only 24 of 53 microarray slides available for analysis. Currently, new experiments are performed for the most rejected samples, although due to small size of biopsy material it will not be always possible to repeat microarray experiments for all contaminated pre-treatment samples. When available, new data will be combined with the data described in this study and the data mining process will be repeated for the full data set. With the increase number of pre- and post- treatment pairs for the same patients it will be possible to create new data sets that will allow eliminate inter patient variations.

Our study shows that from 558 human genes reported in TRANSFAC database to have a transcription factor, 37% have only one associated transcription factor. With majority of genes being regulated by multiple transcription factors, the future of microarray data processing is for pattern recognition methods that allow assignment of a gene to multiple expression patterns compared to clustering algorithms.

TRANSFAC database is regularly updated with more annotations, giving more prior information. Therefore, using modified Bayesian Decomposition with included prior knowledge about groups of co-expressed genes will be even more favorable. On the other hand, enhancement analysis of gene groups received by a pattern

recognition algorithm will receive more power to draw conclusions about transcription factor activity linked to a certain pattern of expression.

Finally, constantly growing information about genes' transcription factors results in increase of information to interpret at final stages of analysis in order to infer activities of upstream signalling pathways. With tens and hundreds of downstream indicators of activity of signalling pathways, it is apparent that such an analysis needs additional tools to make it possible.

6.3 Global Picture

Designed data mining process presented in this study allows recovering underlying expression patterns from observed expression profiles and group genes together. Biological processes can be mapped to these patterns by performing gene group enhancement analysis or directly from the expression pattern behavior. Such interpretation is especially effective, since BD accounts for ability of one gene to be involved in multiple biological processes. Also, inclusion of prior knowledge into the analysis contributes greatly to the strength of the algorithm.

Determining if a biological process activated or deactivated is a very important for analysis of microarray data because of inability to receive direct measurements of genes activity. That is, mRNA levels, which microarrays measure, provide only downstream indications of cell processes, but can be used to determine mechanisms of regulation and activations of signalling pathways that resulted in observed mRNA expression data. The problem of protein activity should be solved in the future with the further development of proteomics, but microarrays are presently the only truly

global measurements tool in functional genomics and reported data analysis process, which includes modified Bayesian Decomposition and Automated Sequence Annotation Pipeline, provides a great tool for analysis of microarray data.

LIST OF REFERENCES

1. Rous, P. and G. John, *Conditional neoplasms and subthreshold neoplastic states: a study of the tar tumors of rabbits.* J Exp Med, 1941. **73**: p. 365-95.

2. zur Hausen, H., *Viruses in human cancers.* Science, 1991. **254**(5035): p. 1167-73.

3. Vineis, P. and K. Husgafvel-Pursiainen, *Air pollution and cancer: biomarker studies in human populations.* Carcinogenesis, 2005. **26**(11): p. 1846-55.

4. Hecht, S.S., *Tobacco carcinogens, their biomarkers and tobacco-induced cancer.* Nat Rev Cancer, 2003. **3**(10): p. 733-44.

5. Saladi, R.N. and A.N. Persaud, *The causes of skin cancer: a comprehensive review.* Drugs Today (Barc), 2005. **41**(1): p. 37-53.

6. Poirier, M.C., *Chemical-induced DNA damage and human cancer risk.* Nat Rev Cancer, 2004. **4**(8): p. 630-7.

7. Golub, T.R., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**(5439): p. 531-7.

8. Yeang, C.H., et al., *Molecular classification of multiple tumor types.* Bioinformatics, 2001. **17 Suppl 1**: p. S316-22.

9. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-52.

10. Garber, M.E., et al., *Diversity of gene expression in adenocarcinoma of the lung.* Proc Natl Acad Sci U S A, 2001. **98**(24): p. 13784-9.

11. Hahn, W.C. and R.A. Weinberg, *Modelling the molecular circuitry of cancer.* Nat Rev Cancer, 2002. **2**(5): p. 331-41.

12. Sjoblom, T., et al., *The consensus coding sequences of human breast and colorectal cancers.* Science, 2006. **314**(5797): p. 268-74.

13. Pearson, G., et al., *Mitogen-activated protein (MAP) kinase pathways: regulation and physiological functions.* Endocr Rev, 2001. **22**(2): p. 153-83.

14. Robinson, M.J. and M.H. Cobb, *Mitogen-activated protein kinase pathways.* Curr Opin Cell Biol, 1997. **9**(2): p. 180-6.

15. Johnson, G.L. and R. Lapadat, *Mitogen-activated protein kinase pathways mediated by ERK, JNK, and p38 protein kinases.* Science, 2002. **298**(5600): p. 1911-2.

16. Chang, L. and M. Karin, *Mammalian MAP kinase signalling cascades.* Nature, 2001. **410**(6824): p. 37-40.

17. Moelling, K., et al., *Regulation of Raf-Akt Cross-talk.* J Biol Chem, 2002. **277**(34): p. 31099-106.

18. Bos, J.L., *ras oncogenes in human cancer: a review.* Cancer Res, 1989. **49**(17): p. 4682-9.

19. Calo, V., et al., *STAT proteins: from normal control of cellular events to tumorigenesis.* J Cell Physiol, 2003. **197**(2): p. 157-68.

20. Libermann, T.A., et al., *Amplification, enhanced expression and possible rearrangement of EGF receptor gene in primary human brain tumours of glial origin.* Nature, 1985. **313**(5998): p. 144-7.

21. Sherr, C.J. and F. McCormick, *The RB and p53 pathways in cancer.* Cancer Cell, 2002. **2**(2): p. 103-12.

22. Steelman, L.S., et al., *JAK/STAT, Raf/MEK/ERK, PI3K/Akt and BCR-ABL in cell cycle progression and leukemogenesis.* Leukemia, 2004. **18**(2): p. 189-218.

23. Kraus, M.H., et al., *Overexpression of the EGF receptor-related proto-oncogene erbB-2 in human mammary tumor cell lines by different molecular mechanisms.* Embo J, 1987. **6**(3): p. 605-10.

24. Marshall, C.J., *Ras effectors.* Curr Opin Cell Biol, 1996. **8**(2): p. 197-204.

25. Lockhart, D.J., et al., *Expression monitoring by hybridization to high-density oligonucleotide arrays.* Nat Biotechnol, 1996. **14**(13): p. 1675-80.

26. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.

27. Griffin, T.J., et al., *Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae.* Mol Cell Proteomics, 2002. **1**(4): p. 323-33.

28. Gygi, S.P., et al., *Correlation between protein and mRNA abundance in yeast.* Mol Cell Biol, 1999. **19**(3): p. 1720-30.

29. Fletcher, C.D., et al., *Diagnosis of gastrointestinal stromal tumors: A consensus approach.* Hum Pathol, 2002. **33**(5): p. 459-65.

30. DeMatteo, R.P., et al., *Two hundred gastrointestinal stromal tumors: recurrence patterns and prognostic factors for survival.* Ann Surg, 2000. **231**(1): p. 51-8.

31. Graadt van Roggen, J.F., M.L. van Velthuysen, and P.C. Hogendoorn, *The histopathological differential diagnosis of gastrointestinal stromal tumours.* J Clin Pathol, 2001. **54**(2): p. 96-102.

32. Kindblom, L.G., et al., *Gastrointestinal pacemaker cell tumor (GIPACT): gastrointestinal stromal tumors show phenotypic characteristics of the interstitial cells of Cajal.* Am J Pathol, 1998. **152**(5): p. 1259-69.

33. Hirota, S., et al., *Gain-of-function mutations of c-kit in human gastrointestinal stromal tumors.* Science, 1998. **279**(5350): p. 577-80.

34. Rubin, B.P., et al., *KIT activation is a ubiquitous feature of gastrointestinal stromal tumors.* Cancer Res, 2001. **61**(22): p. 8118-21.

35. Hirota, S., et al., *Gain-of-function mutations of platelet-derived growth factor receptor alpha gene in gastrointestinal stromal tumors.* Gastroenterology, 2003. **125**(3): p. 660-7.

36. Taylor, M.L. and D.D. Metcalfe, *Kit signal transduction.* Hematol Oncol Clin North Am, 2000. **14**(3): p. 517-35.

37. Druker, B.J., et al., *Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells.* Nat Med, 1996. **2**(5): p. 561-6.

38. Heinrich, M.C., et al., *PDGFRA activating mutations in gastrointestinal stromal tumors.* Science, 2003. **299**(5607): p. 708-10.

39. Heinrich, M.C., et al., *Inhibition of c-kit receptor tyrosine kinase activity by STI 571, a selective tyrosine kinase inhibitor.* Blood, 2000. **96**(3): p. 925-32.

40. Longley, B.J., M.J. Reguera, and Y. Ma, *Classes of c-KIT activating mutations: proposed mechanisms of action and implications for disease classification and therapy.* Leuk Res, 2001. **25**(7): p. 571-6.

41. Demetri, G.D., et al., *Efficacy and safety of imatinib mesylate in advanced gastrointestinal stromal tumors.* N Engl J Med, 2002. **347**(7): p. 472-80.

42. Antonescu, C.R., et al., *Acquired resistance to imatinib in gastrointestinal stromal tumor occurs through secondary gene mutation.* Clin Cancer Res, 2005. **11**(11): p. 4182-90.

43. Tamborini, E., et al., *A new mutation in the KIT ATP pocket causes acquired resistance to imatinib in a gastrointestinal stromal tumor patient.* Gastroenterology, 2004. **127**(1): p. 294-9.

44. De Giorgi, U. and J. Verweij, *Imatinib and gastrointestinal stromal tumors: Where do we go from here?* Mol Cancer Ther, 2005. **4**(3): p. 495-501.

45. Moloshok, T.D., et al., *Application of Bayesian decomposition for analysing microarray data.* Bioinformatics, 2002. **18**(4): p. 566-75.

46. Drmanac, S. and R. Drmanac, *Processing of cDNA and genomic kilobase-size clones for massive screening, mapping and sequencing by hybridization.* Biotechniques, 1994. **17**(2): p. 328-9, 332-6.

47. Schena, M., et al., *Parallel human genome analysis: microarray-based expression monitoring of 1000 genes.* Proc Natl Acad Sci U S A, 1996. **93**(20): p. 10614-9.

48. Shalon, D., S.J. Smith, and P.O. Brown, *A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization.* Genome Res, 1996. **6**(7): p. 639-45.

49. Patterson, T.A., et al., *Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project.* Nat Biotechnol, 2006. **24**(9): p. 1140-50.

50. Halgren, R.G., et al., *Assessment of clone identity and sequence fidelity for 1189 IMAGE cDNA clones.* Nucleic Acids Res, 2001. **29**(2): p. 582-8.

51. Harbig, J., R. Sprinkle, and S.A. Enkemann, *A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array.* Nucleic Acids Res, 2005. **33**(3): p. e31.

52. Fielden, M.R., et al., *GP3: GenePix post-processing program for automated analysis of raw microarray data.* Bioinformatics, 2002. **18**(5): p. 771-3.

53. Medigue, C., et al., *Imagene: an integrated computer environment for sequence annotation and analysis.* Bioinformatics, 1999. **15**(1): p. 2-15.

54. Saeed, A.I., et al., *TM4: a free, open-source system for microarray data management and analysis.* Biotechniques, 2003. **34**(2): p. 374-8.

55. Bidaut, G., et al., *WaveRead: automatic measurement of relative gene expression levels from microarrays using wavelet analysis.* J Biomed Inform, 2006. **39**(4): p. 379-88.

56. Bengtsson, A. and H. Bengtsson, *Microarray image analysis: background estimation using quantile and morphological filters.* BMC Bioinformatics, 2006. **7**: p. 96.

57. Quackenbush, J., *Microarray data normalization and transformation.* Nat Genet, 2002. **32 Suppl**: p. 496-501.

58. Tseng, G.C., et al., *Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects.* Nucleic Acids Res, 2001. **29**(12): p. 2549-57.

59. Yang, Y.H., et al., *Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.* Nucleic Acids Res, 2002. **30**(4): p. e15.

60. Cleveland, W.S. and S.J. Devlin, *Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting.* Journal of the American Statistical Association, 1988. **83**: p. 596-610.

61. Baird, D., P. Johnstone, and T. Wilson, *Normalization of microarray data using a spatial mixed model analysis which includes splines.* Bioinformatics, 2004. **20**(17): p. 3196-205.

62. Chua, S.W., et al., *A novel normalization method for effective removal of systematic variation in microarray data.* Nucleic Acids Res, 2006. **34**(5): p. e38.

63. Wang, J., J.Z. Ma, and M.D. Li, *Normalization of cDNA microarray data using wavelet regressions.* Comb Chem High Throughput Screen, 2004. **7**(8): p. 783-91.

64. Yoon, D., et al., *Two-stage normalization using background intensities in cDNA microarray data.* BMC Bioinformatics, 2004. **5**: p. 97.

65. Berger, J.A., et al., *Optimized LOWESS normalization parameter selection for DNA microarray data.* BMC Bioinformatics, 2004. **5**: p. 194.

66. Wang, D., et al., *A robust two-way semi-linear model for normalization of cDNA microarray data.* BMC Bioinformatics, 2005. **6**: p. 14.

67. Fujita, A., et al., *Evaluating different methods of microarray data normalization.* BMC Bioinformatics, 2006. **7**(1): p. 469.

68. Affymetrix, *Statistical Algorithms Description Document.* Technical Report 701137, 2002. **Rev. 3**.

69. Li, C. and W.H. Wong, *Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.* Proc Natl Acad Sci U S A, 2001. **98**(1): p. 31-6.

70. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-64.

71. Wu, Z., R. Irizarry, and R. Gentlrman, *A model based background adjustment for oligonucleotide expression arrays.* J Am Stat Assoc, 2004. **99**: p. 909-917.

72. Budhraja, V., et al., *Incorporation of gene-specific variability improves expression analysis using high-density DNA microarrays.* BMC Biol, 2003. **1**: p. 1.

73. Hsiao, A., et al., *Variance-modeled posterior inference of microarray data: detecting gene-expression changes in 3T3-L1 adipocytes.* Bioinformatics, 2004. **20**(17): p. 3108-27.

74. Miller, R.A., A. Galecki, and R.J. Shmookler-Reis, *Interpretation, design, and analysis of gene array expression experiments.* J Gerontol A Biol Sci Med Sci, 2001. **56**(2): p. B52-7.

75. Allison, D.B., et al., *Microarray data analysis: from disarray to consolidation and consensus.* Nat Rev Genet, 2006. **7**(1): p. 55-65.

76. Kerr, M.K., M. Martin, and G.A. Churchill, *Analysis of variance for gene expression microarray data.* J Comput Biol, 2000. **7**(6): p. 819-37.

77. Churchill, G.A., *Using ANOVA to analyze microarray data.* Biotechniques, 2004. **37**(2): p. 173-5, 177.

78. Pavlidis, P., *Using ANOVA for gene selection from microarray studies of the nervous system.* Methods, 2003. **31**(4): p. 282-9.

79. Gossett, W.S., *The probable error of a mean.* Biometrika, 1908. **6**: p. 1-25.

80. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.

81. Shedden, K., *Confidence levels for the comparison of microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3**(1): p. Article32.

82. Yang, H. and G. Churchill, *Estimating p-values in small microarray experiments.* Bioinformatics, 2006.

83. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.

84. Heyer, L.J., S. Kruglyak, and S. Yooseph, *Exploring expression data: identification and analysis of coexpressed genes.* Genome Res, 1999. **9**(11): p. 1106-15.

85. Tibshirani, R., et al., *Clustering Methods for the Analysis of DNA Microarray Data.* Stanford University, 1999.

86. Tamayo, P., et al., *Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.* Proc Natl Acad Sci U S A, 1999. **96**(6): p. 2907-12.

87. Brazma, A. and J. Vilo, *Gene expression data analysis.* FEBS Lett, 2000. **480**(1): p. 17-24.

88. Datta, S. and S. Datta, *Comparisons and validation of statistical clustering techniques for microarray gene expression data.* Bioinformatics, 2003. **19**(4): p. 459-66.

89. Alon, U., et al., *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.* Proc Natl Acad Sci U S A, 1999. **96**(12): p. 6745-50.

90. Cho, R.J., et al., *A genome-wide transcriptional analysis of the mitotic cell cycle.* Mol Cell, 1998. **2**(1): p. 65-73.

91. Chu, S., et al., *The transcriptional program of sporulation in budding yeast.* Science, 1998. **282**(5389): p. 699-705.

92. Gasch, A.P., et al., *Genomic expression programs in the response of yeast cells to environmental changes.* Mol Biol Cell, 2000. **11**(12): p. 4241-57.

93. Morgan, B.J. and A.P. Ray, *Non-uniqueness and Inversions in Cluster Analysis.* Appl. Stat., 1995. **44**(1): p. 117-134.

94. Cherepinsky, V., et al., *Shrinkage-based similarity metric for cluster analysis of microarray data.* Proc Natl Acad Sci U S A, 2003. **100**(17): p. 9668-73.

95. Tavazoie, S., et al., *Systematic determination of genetic network architecture.* Nat Genet, 1999. **22**(3): p. 281-5.

96. De Smet, F., et al., *Adaptive quality-based clustering of gene expression profiles.* Bioinformatics, 2002. **18**(5): p. 735-46.

97. Baker, T.K., et al., *Temporal gene expression analysis of monolayer cultured rat hepatocytes.* Chem Res Toxicol, 2001. **14**(9): p. 1218-31.

98. Toronen, P., et al., *Analysis of gene expression data using self-organizing maps.* FEBS Lett, 1999. **451**(2): p. 142-6.

99. Ellestad, L.E., et al., *Gene expression profiling during cellular differentiation in the embryonic pituitary gland using cDNA microarrays.* Physiol Genomics, 2006. **25**(3): p. 414-25.

100. Alsaker, K.V. and E.T. Papoutsakis, *Transcriptional program of early sporulation and stationary-phase events in Clostridium acetobutylicum.* J Bacteriol, 2005. **187**(20): p. 7103-18.

101. Roberts, C.J., et al., *Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles.* Science, 2000. **287**(5454): p. 873-80.

102. Lukashin, A.V. and R. Fuchs, *Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters.* Bioinformatics, 2001. **17**(5): p. 405-14.

103. Gasch, A.P. and M.B. Eisen, *Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering.* Genome Biol, 2002. **3**(11): p. RESEARCH0059.

104. Nam, D., et al., *ADGO: analysis of differentially expressed gene sets using composite GO annotation.* Bioinformatics, 2006. **22**(18): p. 2249-53.

105. Lee, J.S., G. Katari, and R. Sachidanandam, *GObar: a gene ontology based analysis and visualization tool for gene sets.* BMC Bioinformatics, 2005. **6**: p. 189.

106. Busold, C.H., et al., *Integration of GO annotations in Correspondence Analysis: facilitating the interpretation of microarray data.* Bioinformatics, 2005. **21**(10): p. 2424-9.

107. Yu, X., et al., *A system-based approach to interpret dose- and time-dependent microarray data: quantitative integration of gene ontology analysis for risk assessment.* Toxicol Sci, 2006. **92**(2): p. 560-77.

108. Brown, M.P., et al., *Knowledge-based analysis of microarray gene expression data by using support vector machines.* Proc Natl Acad Sci U S A, 2000. **97**(1): p. 262-7.

109. Furey, T.S., et al., *Support vector machine classification and validation of cancer tissue samples using microarray expression data.* Bioinformatics, 2000. **16**(10): p. 906-14.

110. Khan, J., et al., *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.* Nat Med, 2001. **7**(6): p. 673-9.

111. Tatusova, T.A., I. Karsch-Mizrachi, and J.A. Ostell, *Complete genomes in WWW Entrez: data representation and analysis.* Bioinformatics, 1999. **15**(7-8): p. 536-43.

112. Boeckmann, B., et al., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.* Nucleic Acids Res, 2003. **31**(1): p. 365-70.

113. Hubbard, T., et al., *Ensembl 2005.* Nucleic Acids Res, 2005. **33**(Database issue): p. D447-53.

114. Kanehisa, M., *The KEGG database.* Novartis Found Symp, 2002. **247**: p. 91-101; discussion 101-3, 119-28, 244-52.

115. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.

116. Diehn, M., et al., *SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data.* Nucleic Acids Res, 2003. **31**(1): p. 219-23.

117. Lenhard, B., W.S. Hayes, and W.W. Wasserman, *GeneLynx: a gene-centric portal to the human genome.* Genome Res, 2001. **11**(12): p. 2151-7.

118. Rebhan, M., et al., *GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support.* Bioinformatics, 1998. **14**(8): p. 656-64.

119. Alter, O., P.O. Brown, and D. Botstein, *Singular value decomposition for genome-wide expression data processing and modeling.* Proc Natl Acad Sci U S A, 2000. **97**(18): p. 10101-6.

120. Spellman, P.T., et al., *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.* Mol Biol Cell, 1998. **9**(12): p. 3273-97.

121. Bleharski, J.R., et al., *Use of genetic profiling in leprosy to discriminate clinical forms of the disease.* Science, 2003. **301**(5639): p. 1527-30.

122. Mao, R., et al., *Global up-regulation of chromosome 21 gene expression in the developing Down syndrome brain.* Genomics, 2003. **81**(5): p. 457-67.

123. Holter, N.S., et al., *Fundamental patterns underlying gene expression profiles: simplicity from complexity.* Proc Natl Acad Sci U S A, 2000. **97**(15): p. 8409-14.

124. Ghosh, D., *Singular value decomposition regression models for classification of tumors from microarray experiments.* Pac Symp Biocomput, 2002: p. 18-29.

125. Misra, J., et al., *Interactive exploration of microarray gene expression patterns in a reduced dimensional space.* Genome Res, 2002. **12**(7): p. 1112-20.

126. Quackenbush, J., *Computational analysis of microarray data.* Nat Rev Genet, 2001. **2**(6): p. 418-27.

127. Wall, M.E., A. Rechtsteiner, and L.M. Rocha, *Singular value decomposition and principal component analysis*, in *A Practical Approach to Microarray Data Analysis*, D.P. Berrar, W. Dubitzky, and M. Granzow, Editors. 2003, Kluwer: Norwell, MA. p. 91-109.

128. Sanguinetti, G., et al., *Accounting for probe-level noise in principal component analysis of microarray data.* Bioinformatics, 2005. **21**(19): p. 3748-54.

129. Liebermeister, W., *Linear modes of gene expression determined by independent component analysis.* Bioinformatics, 2002. **18**(1): p. 51-60.

130. Alizadeh, A.A., et al., *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.* Nature, 2000. **403**(6769): p. 503-11.

131. Martoglio, A.M., et al., *A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer.* Bioinformatics, 2002. **18**(12): p. 1617-24.

132. Saidi, S.A., et al., *Independent component analysis of microarray data in the study of endometrial cancer.* Oncogene, 2004. **23**(39): p. 6677-83.

133. Zhang, X.W., et al., *Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis.* Eur J Hum Genet, 2005. **13**(12): p. 1303-11.

134. Hyvärinen, A., J. Karhunen, and E. Oja, *Independent Component Analysis*. 2001, New York: John Wiley & Sons.

135. Lee, S.I. and S. Batzoglou, *Application of independent component analysis to microarrays.* Genome Biol, 2003. **4**(11): p. R76.

136. Kim, S.K., et al., *A gene expression map for Caenorhabditis elegans.* Science, 2001. **293**(5537): p. 2087-92.

137. Hsiao, L.L., et al., *A compendium of gene expression in normal human tissues.* Physiol Genomics, 2001. **7**(2): p. 97-104.

138. Lee, D.D. and H.S. Seung, *Learning the parts of objects by non-negative matrix factorization.* Nature, 1999. **401**(6755): p. 788-91.

139. Kim, P.M. and B. Tidor, *Subsystem identification through dimensionality reduction of large-scale gene expression data.* Genome Res, 2003. **13**(7): p. 1706-18.

140. Inamura, K., et al., *Two subclasses of lung squamous cell carcinoma with different gene expression profiles and prognosis identified by hierarchical clustering and non-negative matrix factorization.* Oncogene, 2005. **24**(47): p. 7105-13.

141. Brunet, J.P., et al., *Metagenes and molecular pattern discovery using matrix factorization.* Proc Natl Acad Sci U S A, 2004. **101**(12): p. 4164-9.

142. Fogel, P., et al., *Inferential, robust non-negative matrix factorization analysis of microarray data.* Bioinformatics, 2006.

143. Carmona-Saez, P., et al., *Biclustering of gene expression data by Non-smooth Non-negative Matrix Factorization.* BMC Bioinformatics, 2006. **7**: p. 78.

144. Wang, G., A.V. Kossenkov, and M.F. Ochs, *LS-NMF: a modified non-negative matrix factorization algorithm utilizing uncertainty estimates.* BMC Bioinformatics, 2006. **7**: p. 175.

145. Ochs, M.F., et al., *A new method for spectral decomposition using a bilinear Bayesian approach.* J. Magn. Reson., 1999. **137**: p. 161-176.

146. Bidaut, G., et al., *Bayesian Decomposition analysis of gene expression in yeast deletion mutants.* Methods of Microarray Data Analysis II, 2002: p. 105-122.

147. Moloshok, T.D., et al., *Bayesian Decomposition classification of the project normal data set.* Methods of Microarray Data Analysis III, 2003: p. 211-232.

148. Kossenkov, A.V., G. Bidaut, and M.F. Ochs, *Genes associated with prognosis in adenocarcinomas across studies at multiple institutions.* Methods of Microarray Data Analysis IV, 2005: p. 239-253.

149. Peterson, A.J., A.V. Kossenkov, and M.F. Ochs, *Linking gene expression patterns and transcriptional regulation in Plasmodium falciparum.* Methods of Microarray Data Analysis V, In Press.

150. Besag, J., et al., *Bayesian computation and stochastic systems.* Statistical Science, 1995. **10**: p. 3-41.

151. Geman, S. and D. Geman, *Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images.* IEEE Trans. Pattern Anal. Mach. Intell., 1984. **6**: p. 721-741.

152. Kirkpatrick, S., D. Gelatt Jr, and M.P. Vecchi, *Optimization by simmulated annealing.* Science, 1983. **220**: p. 671-680.

153. Hughes, T.R., et al., *Functional discovery via a compendium of expression profiles.* Cell, 2000. **102**(1): p. 109-26.

154. Pepe, M.S., *Three approaches to regression analysis of receiver operating characteristic curves for continuous test results.* Biometrics, 1998. **54**(1): p. 124-35.

155. Ideker, T., et al., *Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data.* J Comput Biol, 2000. **7**(6): p. 805-17.

156. Rocke, D.M. and B. Durbin, *A model for measurement error for gene expression arrays.* J Comput Biol, 2001. **8**(6): p. 557-69.

157. Lu, X., M. Hauskrecht, and R.S. Day, *Modeling cellular processes with variational Bayesian cooperative vector quantizer.* Pac Symp Biocomput, 2004: p. 533-44.

158. Metz, C.E., *Basic principles of ROC analysis.* Semin Nucl Med, 1978. **8**(4): p. 283-98.

159. Kossenkov, A., et al., *ASAP: automated sequence annotation pipeline for web-based updating of sequence information with a local dynamic database.* Bioinformatics, 2003. **19**(5): p. 675-6.

160. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2006. **34**(Database issue): p. D173-80.

161. Wingender, E., *TRANSFAC, TRANSPATH and CYTOMER as starting points for an ontology of regulatory networks.* In Silico Biol, 2004. **4**(1): p. 55-61.

162. http://www.chem.agilent.com/scripts/generic.asp?lpage=5175.

163. http://www.affymetrix.com/support/index.affx.

164. Chomczynski, P. and N. Sacchi, *Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction.* Anal Biochem, 1987. **162**(1): p. 156-9.

165. Novoradovskaya, N., et al., *Universal Reference RNA as a standard for microarray experiments.* BMC Genomics, 2004. **5**(1): p. 20.

166. Rajotte, D., et al., *Contribution of both STAT and SRF/TCF to c-fos promoter activation by granulocyte-macrophage colony-stimulating factor.* Blood, 1996. **88**(8): p. 2906-16.

167. Narvaez, C.J. and J. Welsh, *Role of mitochondria and caspases in vitamin D-mediated apoptosis of MCF-7 breast cancer cells.* J Biol Chem, 2001. **276**(12): p. 9101-7.

168. Welsh, J., et al., *Impact of the Vitamin D3 receptor on growth-regulatory pathways in mammary gland and breast cancer.* J Steroid Biochem Mol Biol, 2002. **83**(1-5): p. 85-92.

169. Amati, B., et al., *Transcriptional activation by the human c-Myc oncoprotein in yeast requires interaction with Max.* Nature, 1992. **359**(6394): p. 423-6.

170. Kretzner, L., E.M. Blackwood, and R.N. Eisenman, *Myc and Max proteins possess distinct transcriptional activities.* Nature, 1992. **359**(6394): p. 426-9.

171. Nakao, K., et al., *Involvement of an AFP1-binding site in cell-specific transcription of the pre-S1 region of the human hepatitis B virus surface antigen gene.* Nucleic Acids Res, 1989. **17**(23): p. 9833-42.

172. Sawadaishi, K., T. Morinaga, and T. Tamaoki, *Interaction of a hepatoma-specific nuclear factor with transcription-regulatory sequences of the human alpha-fetoprotein and albumin genes.* Mol Cell Biol, 1988. **8**(12): p. 5179-87.

173. Anastasiadis, A.G., et al., *Loss of function of the tissue specific transcription factor HNF1 alpha in renal cell carcinoma and clinical prognosis.* Anticancer Res, 1999. **19**(3A): p. 2105-10.

VITA

PERSONAL INFORMATION

Full name: Andrei Vladimirovich Kossenkov
Place and year of birth: Estonia (USSR), 1979
Country of citizenship: Russian Federation

EDUCATION

Jan 2004 - Jun 2007 PhD in Biomedical Science (Bioinformatics) at the Drexel University. Philadelphia, PA.

Sep 2001 - Feb 2003 Master of Science in Applied Mathematics (Bioinformatics) at the Moscow Engineering Physics Institute (State University). Moscow, Russia.

Sep 1997 - Feb 2001 Bachelor of Science in Applied Mathematics at the Moscow Engineering Physics Institute (State University). Moscow, Russia.

SELECTED PUBLICATIONS

Wang, G., **Kossenkov, A.V.**, Ochs, M.F.: LS-NMF: a modified non-negative matrix factorization algorithm utilizing uncertainty estimates. *BMC Bioinformatics*, 2006

A. Kossenkov, G. Bidaut, and M.F. Ochs: Genes associated with prognosis in adenocarcinomas across studies at multiple institutions, in *J. Shoemaker and S. Lin, Editors, Methods of Microarray Data Analysis IV*, 239-253., Kluwer Academic, Boston, 2005.

Kossenkov, A., Manion, F.J., Korotkov, E., Moloshok, T.D., Ochs, M.F.: ASAP: automated sequence annotation pipeline for web-based updating of sequence information with a local dynamic database. *Bioinformatics* 19(5): 675-6, 2003